PHOTOGRAPH THIS SHEET

AD-A189 604

DTIC ACCESSION NUMBER

LEVEL

INVENTORY 1

AFWAL-TR-87-1148

DOCUMENT IDENTIFICATION

29 JAN 88

[redacted] has been approved for public release and sale; its [redacted] is unlimited.

DISTRIBUTION STATEMENT

ACCESSION FOR	
NTIS GRA&I	X
DTIC TAB	☐
UNANNOUNCED	☐
JUSTIFICATION	
BY	
DISTRIBUTION /	
AVAILABILITY CODES	
DIST	AVAIL AND/OR SPECIAL
A-1	

DISTRIBUTION STAMP

DTIC
ELECTE
S MAR 04 1988 D
E

DATE ACCESSIONED

DTIC COPY INSPECTED

DATE RETURNED

88 3 04 039

DATE RECEIVED IN DTIC

REGISTERED OR CERTIFIED NO.

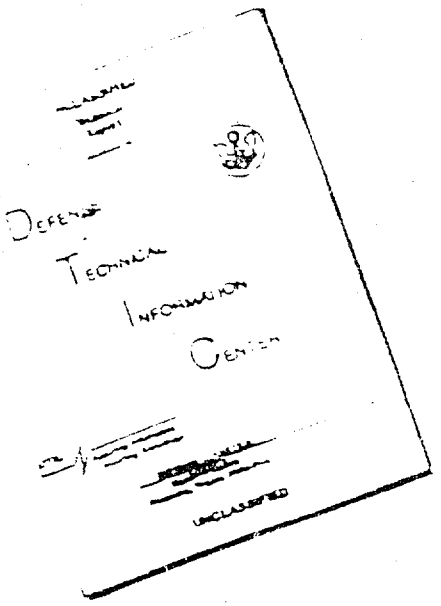
PHOTOGRAPH THIS SHEET AND RETURN TO DTIC-FDAC

DTIC FORM 70A
MAR 86

DOCUMENT PROCESSING SHEET

PREVIOUS EDITION MAY BE USED UNTIL STOCK IS EXHAUSTED.

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

AFWAL-TR-87-1148

REDUCED TOLERANCE IMAGING II

J.R. FIENUP, et al

Environmental Research Institute of Michigan
Advanced Concepts Division
P.O. Box 8618
Ann Arbor, MI 48107-8618

JANUARY 1988

Final Report, Vol. II for Period October 1984-October 1986

Avionics Laboratory
Air Force Wright Aeronautical Laboratories
Air Force Systems Command
Wright Patterson Air Force Base, OH 45433-6543

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS (none)
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4 PERFORMING ORGANIZATION REPORT NUMBER(S) 167400-101-F	5. MONITORING ORGANIZATION REPORT NUMBER(S) AFWAL-TR-87-1148
6a. NAME OF PERFORMING ORGANIZATION Environmental Research Institute of Michigan 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION Avionics Laboratory (AFWAL/AARI) AF Wright Aeronautical Laboratories
6c. ADDRESS (City, State, and ZIP Code) P.O. Box 8618 Ann Arbor, MI 48107	7b. ADDRESS (City, State, and ZIP Code) Wright-Patterson AFB, Ohio 45433-6543
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Defense Advanced Research Projects Agency 8b. OFFICE SYMBOL (If applicable) TTO	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F33615-83-C-1046, DARPA Order 5205
8c. ADDRESS (City, State, and ZIP Code) 1400 Wilson Blvd. Arlington, VA 22209	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 61101F; PROJECT NO ILIR; TASK NO. 83; WORK UNIT ACCESSION NO 05
11. TITLE (Include Security Classification) Reduced Tolerance Imaging II	
12. PERSONAL AUTHOR(S) J.R. Fienup, J.N. Cederquist, T.R. Crimmins, R.G. Paxman, D.L. Neuhoff and G. Goldstein	
13a. TYPE OF REPORT (Vol. II) Final Technical 13b. TIME COVERED FROM 10/84 TO 10/86	14. DATE OF REPORT (Year, Month, Day) 1988, January 29 15. PAGE COUNT 292
16. SUPPLEMENTARY NOTATION This research was partially funded by the inhouse independent research fund.	

17 COSATI CODES		
FIELD	GROUP	SUB-GROUP
20	06	
20	14	

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Phase retrieval
Image reconstruction

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

Reduced tolerance imaging is a concept wherein an imaging system is designed with reduced performance requirements permitting large phase errors in the received signal, and the full performance level is recovered through use of post-detection processing using phase retrieval techniques yielding a diffraction-limited reconstructed image. This report describes a two-year effort to develop reduced-tolerance imaging techniques. An estimation theoretic (Cramer-Rao) lower bound on the error in estimating a coherent image from (1) far-field (Fourier) intensity (squared modulus) measurements and from (2) electromagnetic field measurements with phase errors were derived for the case of additive Gaussian detector noise. Uniqueness of reconstruction from Fourier modulus assuming a support constraint known a priori was proven for a particular class of objects -- sampled objects whose support (the area in which the object has non-zero values) has a convex hull with no parallel sides. A closed-form recursive reconstruction algorithm was developed for reconstructing such objects

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT X UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Mr. William Martin	22b. TELEPHONE (Include Area Code) (513) 255-6361 22c. OFFICE SYMBOL AARI

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This report has been reviewed by the Office of Public Affairs (ASD/PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

This technical report has been reviewed and is approved for publication.

1st Lt Robert Hudson Fetner III
Project Engineer
Electro-Optics Techniques Group
Electro-Optics Branch

Donald L. Tomlinson, Chief
Electro-Optics Techniques Group
Electro-Optics Branch
Mission Avionics Division

FOR THE COMMANDER

GALE D. URBAN, Chief
Electro-Optics Branch
Mission Avionics Division
Avionics Laboratory

If your address has changed, if you wish to be removed from our mailing list, or if the addressee is no longer employed by your organization please notify AFWAL/AARI, Wright-Patterson AFB, OH 45433-6543 to help us maintain a current mailing list.

Copies of this report should not be returned unless return is required by security considerations, contractual obligations, or notice on a specific document.

block # 19 continued

from their autocorrelation functions. Simulations showed the closed-form solution to be very sensitive to noise compared with iterative Fourier transform algorithms, particularly if the corners of the object are dim. The uniqueness proof is important, however, because it is useful for predicting support constraints for which iterative reconstruction is facilitated. Several potential constraints for use in reconstruction algorithms were examined briefly, but support and nonnegativity are the only two constraints that have been extensively exploited. In the case of illumination with an active system, the edges of the illumination pattern are unavoidably tapered which can cause algorithm convergence problems; however, this problem was largely overcome by modification to the iterative transform algorithm using an expanding mask. Successful reconstructions were obtained in the presence of large amounts of edge tapering. Image quality was measured as a function of both edge tapering and noise. Alternative reconstruction algorithms were studied, including various gradient search algorithms (for which analytic expressions for the gradient of the error metric were derived) and a modeling approach, but they have not yet been developed to the point where they out-perform the iterative transform algorithm. A comparison of deterministic and stochastic approaches to reconstruction was also made. A laboratory experiment using active coherent illumination was performed in which a rough object was illuminated by a laser with a known (triangular) illumination pattern and the far-field (Fourier) intensity pattern was detected. Using the iterative transform algorithm employing knowledge of the support of the illumination pattern, an image was reconstructed from the Fourier intensity data. The reconstructed image compared favorably with a conventional "ground truth" image, giving experimental confirmation of the analytical and simulation results. Data was also taken in a passive incoherent experiment in which an object was imaged through a simulated segmented-aperture telescope for which the segments were misaligned. A specific tactical application using 10.6μm laser illumination was analyzed and a system architecture was developed. It was determined that the reduced-tolerance imaging concept would be useful for obtaining finer resolution with a simpler system and with less impact on the platform geometry than would be possible by conventional means.

In summary, the studies performed thus far indicate that the reduced-tolerance imaging concept is a promising approach worthy of continued consideration for reducing size, weight, complexity and cost of fine-resolution imaging systems.

PREFACE

The work reported here was performed in the Optical Science Laboratory of the Advanced Concepts Division, Environmental Research Institute of Michigan (ERIM). The work was sponsored by the Defense Advanced Projects Agency (DARPA) through the Air Force Wright Aeronautical Laboratories (AFWAL) under Contract F33615-83-C-1046, DARPA Order 5205. At AFWAL/AARI, the Project Monitor initially was Lt. Michael Roggemann, and then was Lt. Robert Fetner and the Program Manager was Mr. William Martin.

This final technical report covers work performed from 15 October 1984 to 14 October 1986. This report is cumulative and includes material reported in the Interim Technical Report, "Reduced Tolerance Imaging I," ERIM Report No. 167400-83-T, July 1986, which covered the period 15 October 1984 to 14 October 1985. The principal investigator at ERIM was James R. Fienup. Major contributors to this work were Jack N. Cederquist, Thomas R. Crimmins, John D. Gorman and Richard G. Paxman. Additional contributors were Carl C. Aleksoff, James T. Clinthorne, Greg A. Dale, Gene Goldstein, Jack M. Keoshian, Timothy Klepaczyk, William H. Licata, Ivan J. LaHaie, Joseph C. Marron, David L. Neuhoff, Stanley R. Robinson, Nicola S. Subotic, Anthony M. Tai and Christopher C. Wackerman.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

SECTION 1
INTRODUCTION AND OVERVIEW

1.1 BACKGROUND

In many imaging scenarios that require fine resolution at long ranges, phase errors limit the achievable resolution and prevent diffraction-limited imaging. The phase errors may arise from a variety of sources, including atmospheric turbulence, misaligned or aberrated optics, motion compensation errors, local oscillator errors, and waveform generator errors. The conventional approach for obtaining diffraction-limited imagery is to build increasingly more complex sensor hardware having tight tolerances on its various components to achieve the desired phase stability.

An alternative approach is to build hardware having reduced tolerances on its phase stability, and correct for the phase errors by employing a phase retrieval algorithm in a post-processing stage. In some instances a sensor can be used that is capable of measuring intensity only and does not measure the phase. In either case a phase retrieval algorithm is used to retrieve the phase. This approach is called Reduced Tolerance Imaging (RTI). Using this approach, we can potentially achieve diffraction-limited imaging using a sensor system that is simpler, cheaper, lighter-weight and smaller than a conventional sensor.

In order for a phase retrieval algorithm to work, it is necessary to have some form of a priori information about, or constraints on, the image. Examples of such constraints that have been useful in the past are nonnegativity (applicable to incoherent imaging) and knowledge of the object's support (knowing its width or shape, information which is available for objects on dark backgrounds or if one controls the pattern of radiation that illuminates the object).

Several important issues must be addressed to make the RTI concept feasible. Constraints must be found that are powerful enough to ensure that the retrieved phase and the reconstructed image are uniquely related to the measured data. The relationship between the reconstructed image and the measured data must be robust enough that it is not overly sensitive to noise or other imperfections in the data or constraints. Reconstruction algorithms must be found that converge reliably to a solution with a reasonable amount of computation and in the presence of realistic amounts of noise.

This report describes the results of a 2-year program for initial development of the Reduced Tolerance Imaging concept.

1.2 OVERVIEW OF ACCOMPLISHMENTS

In this section the principal results of the RTI program will be briefly summarized. They are reported in detail in the sections and appendices that follow.

We would like to know how well we could ever hope to reconstruct an image from the given data and constraints. Then we would know whether current reconstruction algorithms are good enough or further development is needed. We would also be able to evaluate and compare various measurement schemes without having to develop reconstruction algorithms for each. This can be done using estimation-theoretic lower bounds on the reconstruction errors. The Cramer-Rao lower bound was derived for the case of far-field intensity measurements with additive Gaussian noise. The lower bound was computed and compared with actual errors experienced in imagery reconstructed from simulated data. These results demonstrate the usefulness of estimation theory for predicting system performance. Cramer-Rao analysis was also performed for a second situation, in which complex-valued measurements of the Fourier transform (fields) are measured, and the measurements are corrupted both by phase

errors and by additive noise. Section 2 and Appendices A and B describe these results.

For discrete, or sampled, objects of a certain type, a closed-form recursive reconstruction algorithm was developed. It reconstructs an image from the autocorrelation function which is obtained by inverse Fourier transforming the measured Fourier intensity data. Unfortunately the closed-form reconstruction algorithm was found to be very sensitive to noise, especially if the corner or vertex points within the object are dim compared with interior points. Although the closed-form reconstruction algorithm is of little practical use because of its sensitivity to noise, it has provided valuable insights into the reconstruction problem. It constitutes a uniqueness proof for the class of objects for which it is applicable and suggests illumination pattern shapes that are advantageous. These results are described in Section 3 and Appendices C and D.

Since image reconstruction with degraded Fourier phase or no Fourier phase requires a priori constraints on the object, it is imperative that object constraints that are sufficiently powerful and robust be found. The vast majority of the work to date has concentrated on two constraints: support, or shape (which occurs naturally for imaging satellites and may be forced by an illumination pattern) and nonnegativity (which occurs naturally for most passive incoherent imaging problems). Issues relating to these and other potential constraints are discussed in Section 4.

When a support constraint is imposed by using an active illumination pattern at the target to achieve the desired known shape, the principal problem is diffraction effects at the edges of the illumination pattern. This makes the illumination pattern fall off slowly and smoothly, i.e., it is tapered, rather than falling off abruptly as would be preferred. Sidelobes of the illumination pattern also prevent it from going exactly

to zero. We have found that reconstruction is much easier when there is little or no tapering of the illumination pattern. Previous versions of the iterative reconstruction algorithm were unsuccessful in reconstructing complex-valued images when large amounts of taper were present. Improved versions of the algorithm, employing an "expanding mask," were developed, and this resulted in a greatly improved result. It consists of using an unrealistically small support constraint for early iterations, which forces the energy of the image to be better centered within the true support constraint, and using progressively larger support constraints for later iterations. It was found that image quality degrades as the amount of taper and of noise increases. With no taper present, the image quality degrades slowly with increasing noise. With tapered illumination, however, the image quality degrades faster. It was also shown that the present algorithms have room for further improvement in terms of achieving better quality imagery in the presence of taper and noise. Section 5 defines the improved algorithm employing the expanding mask, and shows experimental reconstruction results with different types of illumination patterns and for tapered illumination and additive noise.

The iterative algorithm described in Section 5 is one of several possible approaches to solving the phase retrieval problem. Improved algorithms are sought which are faster and more robust. One family of alternative algorithms are the gradient search algorithms. They consist of defining a merit function, computing the gradient of the merit function with respect to a set of parameters, and searching in the parameter space for a minimum of the merit function by moving in the direction of the negative of the gradient (the global minimum of the merit function defines the solution, the reconstructed image). Merit functions that were examined include the amount by which the modulus of the Fourier transform of an object estimate differs from the measured Fourier modulus data and the amount by which an output image violates the object-domain constraints. Parameter spaces that were investigated

include the space of object estimates and the space of Fourier phase estimates. Closed-form expressions for the gradients were derived, and it was shown that the entire gradients can be efficiently computed using a small number of fast Fourier transforms. Gradient search algorithms were tested on the computer with mixed results. They show promise and should be developed further. These results are described in Section 6 and Appendices E, F and G.

Another approach to solving the phase retrieval problem is a modeling approach. The complex Fourier transform or pieces of it are modeled by a parameterized function. The measured Fourier modulus is least-squares fitted to the modulus of the model to determine the unknown parameters. Then the parameters are inserted into the complex model which is evaluated to determine the phase. Attempts thus far to make the modeling approach work were unsuccessful. It is likely that the models used were not appropriate to the complex Fourier transforms of interest. Better models would be needed before further work on this approach should be pursued. This work is discussed in Section 7.

The vast majority of the phase retrieval work prior to the current effort revolved around analysis and computer simulations. Since the computer simulations implicitly assume a discrete model for the object, there is a danger that the real, continuous world might behave differently. For this and other reasons it is very important to demonstrate the feasibility of phase retrieval using real data collected in the laboratory which include the important real-world effects. Two experiments were performed: an active coherent experiment and a passive incoherent experiment. In the active coherent experiment the target was illuminated with a laser beam pattern having the desired illumination shape. A lens formed the far-field (Fourier transform) at a plane where the intensity pattern was detected. A second channel including imaging optics was used to form a "ground truth" image. The far-field detected intensity was corrected for camera response and square-rooted to form an

estimate of the modulus of the Fourier transform of the target field. An image was reconstructed from this data using knowledge of the shape of the illumination pattern. The reconstructed image compares favorably with the object and the ground-truth image. This result demonstrates the feasibility of the phase retrieval approach, at least in a controlled, laboratory environment. In addition a passive incoherent experiment was set up and data was taken. An object was imaged through a simulated segmented-aperture telescope for which the segments were misaligned. However it was not possible to complete the passive experiment under the current funding. Section 8 and Appendix H describe the experimental efforts and results.

A specific imaging application and architecture was analyzed. An active optical system using a 10.6μm laser illuminator would detect reflected intensities over a conformal detector array that would be much larger than the largest aberration-free telescope aperture that could be practically mounted on an airborne platform. Image quality and size dictated the laser power necessary for tactically useful ranges, and the required laser power was within the present state of the art. This is described in Section 9.

For coherent images we can view the phase retrieval problem in a stochastic framework as well as the more conventional deterministic framework. These two approaches are compared in Appendix I.

In summary, the idea of using phase retrieval algorithms with image-domain constraints for reducing the tolerances required on imaging system components continues to look promising. Requirements on transmit power and illumination pattern tapering have been quantified and appear to be achievable for a specific 10.6μm laser-illumination system that was analyzed. Laboratory experimental results, although preliminary, confirm the analytical and computer results. These promising results,

together with the high potential payoff of reduced size, weight, cost, and complexity, suggest that further development of the approach should be undertaken for specific applications.

1.3 Recommendations

It is recommended that the reduced-tolerance imaging concept be further developed along two major directions: application to specific sensors and expanded understanding of the underlying theoretical issues. In this section, research topics relating to two of the most promising tactical applications, an active laser system and a synthetic aperture radar, will be described, and the issues relating to the underlying theory will be discussed.

The use of short-wavelength active imaging systems using laser illumination and a conformal detector array served as the dominant focus for the research reported here and is the best understood of the potential tactical imaging scenarios utilizing phase retrieval. For this application the most important next step is to expand upon the preliminary laboratory experimental results reported in Section 8 by simulating in the laboratory the many real-world effects that are described in Section 8.1.1. Subsequently the technique should be demonstrated at longer ranges in rooftop-based outdoor test-range experiments leading toward an airborne demonstration. Further development of reconstruction algorithms for improved speed and robustness is also required, particularly for the case of larger amounts of tapering of the edges of the illumination pattern.

The second application of the reduced-tolerance imaging concept which should be vigorously pursued is synthetic aperture radar (SAR), particularly as a means for overcoming the lack of adequate motion compensation. Examples of specific applications include rapidly manoeuvering airborne platforms and light-weight platforms lacking

inertial navigation sensors. The SAR applications differ from the laser-illumination systems, emphasized in the research reported here, in some important ways. For SAR we cannot as easily obtain a useful illumination-pattern support constraint, but, on the other hand, we can achieve similar effects by using no-return areas such as lakes, shadows, and smooth runways and roads. The no-return areas in SAR imagery will occupy a much smaller fraction of the total area of the image than the non-illuminated areas in the case of laser-illuminated imagery. This makes image reconstruction more difficult for the case of SAR. However, for SAR we measure the phase which, although it is distorted by motion compensation errors, can be used to start the reconstruction processing with a blurred image, whereas for the laser illumination case it was assumed that there was no phase information at all. Furthermore, the SAR phase errors are primarily one-dimensional (along track) and so the number of unknowns is far less than for the laser illumination case. These factors should make image reconstruction much easier for the case of SAR. This combination of less constraints but more available phase information makes it difficult to predict how well the reconstruction process will work for a realistic SAR scenario based on the initial results shown in this report. Therefore the reduced-tolerance imaging technique should be developed specifically for the types of constraints and data available for SAR scenarios of interest. This would include algorithm development that would make use of no-return areas imbedded in images, use partial phase information that is measured, and take advantage of the one-dimensional nature of the phase error. This should first be done for computer-simulated data and constraints, and later for SAR data collections available at ERIM. An assessment of the types, amounts, and quality of constraints typically available in SAR images of interest would be included in the study. Both analytically and by computer simulation, a prediction of image quality as a function of constraints and data properties should be made. Finally, data collection experiments should be defined and carried out to demonstrate the concept in the field.

Underlying the reduced-tolerance imaging concept are basic phase retrieval algorithms and uniqueness questions that have been investigated by several groups of researchers around the world for over a decade. The most successful algorithm to date is the iterative transform algorithm developed at ERIM that was used in the simulation experiments reported in Sections 2.5, 4.1, and 5. Sections 3, 6 and 7 describe attempts to improve on that algorithm. Further work along these lines is necessary to arrive at algorithms with improved speed and image quality. The gradient search algorithms described in Section 6 appear to be promising and should be developed further. Newton-Raphson and other approaches should also be explored. Continued optimization of and improved variations on the iterative transform algorithm should be pursued, with particular attention paid to characterizing modes of stagnation and developing techniques for overcoming them.

The uniqueness question can be stated as follows: with what confidence can we say that a reconstructed image does not differ substantially from a true image of the object? As reported in Section 3, it was proved that certain illumination patterns guarantee a unique solution for the case of a sampled model of the object; but it is not clear how this carries over to the real world of continuous objects. The answer to this question may come from the abstract mathematical field of analytic functions of several complex variables. However, despite efforts by several workers in the US and England, a comprehensive answer to the uniqueness question remains to be developed. Additional questions, such as how illumination pattern tapering affects the uniqueness of the reconstructed image should also be addressed. To answer these questions, a multifaceted approach should be pursued, including use of the theory of analytic functions of several complex variables, polynomial factorization theory, and brute-force computer searches through the solution space for small images.

With this combination of investigations into specific applications and underlying theory, the reduced tolerance imaging concept can be developed to the point of providing a near-term solution to pressing sensor-system problems.

SECTION 2
INFORMATION THEORETIC LOWER BOUNDS FOR PHASE RETRIEVAL

2.1 INTRODUCTION

In phase retrieval problems, it is desired to estimate the phase of the Fourier transform of an object given measurements of the intensity (i.e., squared modulus) of the Fourier transform. This is equivalent to estimating the object itself because of the Fourier transform relationship. The iterative Fourier transform algorithm has had great success in making such object estimates from Fourier intensity data and object constraint information [2.1, 2.2]. However, other than through empirical results [2.3], it has not been known how the error in the object estimate depends on measurement noise, constraint information, and other parameters describing the problem.

Results in estimation theory include a number of methods whereby lower bounds on the mean-squared error of the object estimate may be calculated. These methods use knowledge of the measurement procedure, the statistics of the object, and the statistics of the noise process to compute an error lower bound. An important feature is that these methods do not require specification of the algorithm used to compute the object estimate from the measured data. The lower bound, then, is independent of the algorithm and therefore indicative of the best possible estimation performance given the chosen measurements and the underlying statistics.

The Cramer-Rao lower bound is the most often used lower bound because it is usually the least difficult to compute. It has been used in a large number of single and multiple parameter and time-varying waveform estimation problems with great success [2.4]. Algorithms exist which produce estimates that achieve the Cramer-Rao bound in problems in

which the measurements are linearly related to the parameters to be estimated, the noise is additive, and the statistics are Gaussian. In nonlinear problems (of which phase retrieval will be an example), the lower bound can usually be approached only at high signal-to-noise ratios [2.4, 2.5]; nonetheless, the lower bound is generally regarded as an important first step in evaluating and designing measurement procedures and parameter estimation algorithms for these problems.

A specific phase retrieval problem is defined in Section 2.2. In Section 2.3, the Cramer-Rao method for computing the error lower bound is reviewed and general notation is developed. In Section 2.4, an explicit expression for the phase retrieval error lower bound is derived. In Section 2.5, a comparison is given of the lower bound to errors found in computer simulations using an iterative Fourier transform phase retrieval algorithm. A second phase retrieval problem is defined in Section 2.6 and an expression for the lower bound given. Conclusions and suggestions for further research are given in Section 2.7. A tutorial review of the Cramer-Rao method is presented in Appendix A. A method for determining Cramer-Rao lower bounds in a wider class of phase retrieval problems and the specific result given in Section 2.6 are derived in Appendix B.

2.2 PHASE RETRIEVAL PROBLEM DEFINITION

From the many combinations of possible phase retrieval problems and underlying assumptions, the following specific example is chosen. It is desired to estimate a two-dimensional, complex-valued object f_m from real-valued measurements S_p where $m = (m_1, m_2)$; $m_1, m_2 = 0, 1, \ldots, M-1$ and $p = (p_1, p_2)$; $p_1, p_2 = 0, 1, \ldots, 2M-1$. The measurements are related to the object by

$$S_p = cI_p + N_p \tag{2-1}$$

and

$$I_p = \left| \sum_m w_m f_m \exp\left[\frac{-i2\pi\langle m, p\rangle}{2M}\right] \right|^2 \qquad (2\text{-}2)$$

where I_p is the intensity (squared modulus) of the discrete Fourier transform of f, c is a proportionality constant, N_p is additive noise, $\langle m, p\rangle = m_1p_1 + m_2p_2$, and summation over m implies the double summation over m_1 and m_2. Object constraint information is essential for estimating the object. The weighting array w_m is explicitly included in Eq. (2-2) to allow arbitrary support constraints to be placed on the object. For an object of M by M resolution elements, Nyquist sampling requires a measurement array of size 2M by 2M because the squared modulus has twice the bandwidth of the complex-valued Fourier transform. It will be convenient later to consider w, f, S, I, and N as vectors. The phase retrieval problem is to estimate the object f given the set of measurements S and knowledge of the constraint that the product w_mf_m is zero wherever w_m is known to be zero.

This mathematical statement can represent a number of applications in which phase retrieval problems arise. For example, consider the measurement geometry shown in Figure 2-1. A known, complex-valued, monochromatic wavefront w(x, y) illuminates an unknown, complex-valued object f(x, y). Alternatively, for the wavefront sensing problem, an unknown monochromatic wavefront f(x, y) may pass through a known aperture having known complex transmittance w(x, y). The intensity I(u, v) in a measurement plane located a distance R from the object plane is [2.6]:

$$I(u,v) = \frac{1}{(\lambda R)^2} \left| \int\int w(x,y) f(x,y) \exp\left[\frac{-i2\pi(ux + vy)}{\lambda R}\right] dxdy \right|^2 \qquad (2\text{-}3)$$

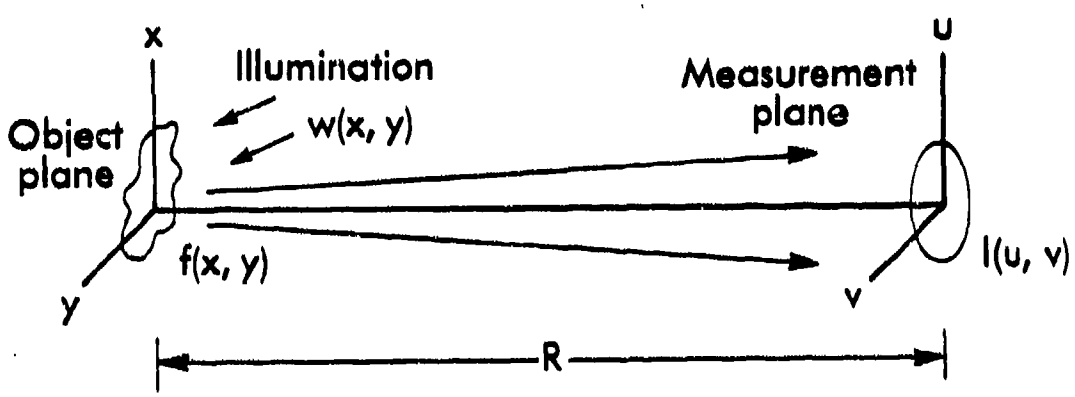


FIGURE 2-1. MEASUREMENT GEOMETRY FOR PHASE RETRIEVAL PROBLEM.

where λ is the wavelength and it is assumed that R is sufficiently great that the Fraunhofer approximation can be made. A discrete set of measurements S is made with

$$S_p = \eta T \int_{\Delta A} I(u, v)\, du\, dv + N_p \qquad (2\text{-}4)$$

where η is the detector efficiency, T is the detector integration time, ΔA is the area of a detector element, N_p is the detector noise, and the subscript $p = (p_1, p_2)$ indexes over the measurement plane. A phase retrieval method (e.g., an iterative Fourier transform algorithm) would be applied to the measurement set S using the object constraint provided by the illumination pattern w to give an estimate of a sampled version of the object f. Conversion of Eqs. (2-3) and (2-4) into discrete form gives, for this application, a value for the constant c in Eq. (2-1) of $\eta T \Delta A (\Delta a/\lambda R)^2$ where Δa is the square area associated with an object sample.

The complex-valued object f can be written in terms of its real and imaginary parts,

$$f = f_m^r + i f_m^i \,. \qquad (2\text{-}5)$$

Equation (2-2) then becomes

$$I_p = \left| \sum_m w_m \left(f_m^r + i f_m^i \right) \exp \left[\frac{-i\pi \langle m,\, p \rangle}{M} \right] \right|^2 \,. \qquad (2\text{-}6)$$

2.3 CRAMER-RAO LOWER BOUND

It can be proven that the variance of any unbiased estimate of a component of a random vector is greater than or equal to the corresponding diagonal element of the inverse of what is called the Fisher information matrix. The value of the diagonal element is the Cramer-Rao lower bound. The elements of the Fisher information matrix depend in turn upon the second partial derivatives of the joint probability distribution of the measurement vector and the vector to be estimated. This result is proven primarily by the use of the Schwarz inequality [2.4].

Application of the Cramer-Rao method for determining lower bounds on estimation errors to a specific problem must therefore begin with a determination of the statistics of the parameters to be estimated and of the noise [2.4, 2.7]. In this analysis, it is assumed that f_m^r, f_m^i, and N_p are each statistically independent, zero mean, Gaussian random variables with variances $(\sigma_f)^2/2$, $(\sigma_f)^2/2$, and $(\sigma_N)^2$ respectively. Note that this implies that f_m is zero mean and that the variance of f_m (i.e., the expected value of $|f_m|^2$) is $(\sigma_f)^2$. Since the Cramer-Rao method can not be directly applied to the complex-valued random variables f_m, this section develops a notation for applying the method to the real-valued f_m^r and f_m^i.

By the definition of conditional probability,

$$p(S, f) = p(S|f)p(f) \tag{2-7}$$

where $p(S, f)$ is the joint probability density of S and f, $p(S|f)$ is the conditional probability density of S given f, and $p(f)$ is the probability density of f. (Recall that f and S are vectors.) The assumption of Gaussian statistics gives

$$p(f) = \prod_m \frac{1}{\pi\sigma_f^2} \exp\left[-\frac{(f_m^r)^2 + (f_m^i)^2}{\sigma_f^2}\right] \tag{2-8}$$

and, using Eqs. (2-1) and (2-6) which imply that $p(S|f) = p(N = S - cI)$,

$$p(S|f) = \prod_p \frac{1}{\sigma_N\sqrt{2\pi}} \exp\left[\frac{-(S_p - cI_p)^2}{2\sigma_N^2}\right] \tag{2-9}$$

where, by Eq. (2-6), I is a function of f.

The Cramer-Rao method continues by defining the Fisher information matrix J in terms of the probability density functions. For the present problem, where it is desired to estimate the statistically independent real and imaginary parts of f, a workable notation is to partition J into four submatrices:

$$J = \left[\begin{array}{c|c} J^{rr} & J^{ri} \\ \hline J^{ir} & J^{ii} \end{array}\right]. \tag{2-10}$$

J is of dimension $2M^2$ by $2M^2$ (representing the M^2 independent f_m^r plus the M^2 independent f_m^i) and each of the submatrices is of dimension M^2 by M^2. The elements of the submatrices are defined by, for example [2.4, 2.7],

$$J_{mn}^{rr} = -E\left[\frac{\partial^2 \ln p(S, f)}{\partial f_m^r \partial f_n^r}\right] \tag{2-11}$$

where $E[\cdot]$ denotes expectation taken over both f and N and the partial derivative holds S constant. The other submatrices are defined by appropriate substitution of the superscripts r and i. It is assumed that these and any other required derivatives exist. This assumption is valid for the phase retrieval problem.

The Cramer-Rao method concludes by determining the inverse J^{-1} of the Fisher information matrix J. The diagonal elements of J^{-1} are the desired lower bound on the mean-squared error of the object estimate $\hat{f}$. From the convention used to define J, the upper left diagonal elements of J^{-1} refer to f_m^r and the lower right elements to f_m^i. If J^{-1} is similarly partitioned into four submatrices:

$$J^{-1} = \left[\begin{array}{c|c} K^{rr} & K^{ri} \\ \hline K^{ir} & K^{ii} \end{array}\right], \tag{2-12}$$

then the Cramer-Rao lower bound $(e_{0m})^2$ on the mean-squared error, $E[|f_m - \hat{f}_m|^2]$, in the estimate $\hat{f}_m$ of object component f_m, is the sum of the diagonal elements for f_m^r and f_m^i:

$$e_{om}^2 = K_{mm}^{rr} + K_{mm}^{ii} . \tag{2-13}$$

This is the quantity which the following analysis seeks. Strictly, the lower bound is only for asymptotically unbiased estimates of f. It is beyond the scope of this work to determine whether particular phase retrieval algorithms give unbiased estimates.

2.4 LOWER BOUND FOR PHASE RETRIEVAL

Substituting Eqs. (2-8) and (2-9) into Eq. (2-11), differentiating, and discarding a term with zero expected value gives [2.4]

$$J_{mn}^{rr} = \frac{c^2}{\sigma_N^2} \sum_p E\left[\frac{\partial I_p}{\partial f_m^r} \frac{\partial I_p}{\partial f_n^r}\right] + \frac{2\delta_{mn}}{\sigma_f^2} \qquad (2\text{-}14)$$

where δ_{mn} is the Kronecker delta function. Similar results hold for the other submatrices of J except that J^{ri} and J^{ir} have no δ_{mn} term. It is important to note that this result holds for any function I of the parameter f. It does not assume that the measurements are of the Fourier intensity.

Equation (2-6) can now be used to compute the first term on the right hand side of Eq. (2-14). First,

$$\frac{\partial I_p}{\partial f_m^r} = w_m^* \sum_j w_j \left[f_j^r + i f_j^i\right] \exp\left[\frac{-i\pi\langle j - m, p\rangle}{M}\right] + \text{c.c.} \qquad (2\text{-}15)$$

Then,

$$\sum_p E\left[\frac{\partial I_p}{\partial f_m^r} \frac{\partial I_p}{\partial f_n^r}\right] =$$

$$\sum_p E\left[w_m^* w_n^* \sum_j \sum_k w_j w_k \left(f_j^r + i f_j^i\right)\left(f_k^r + i f_k^i\right) \exp\left[\frac{-i\pi\langle j + k - m - n, p\rangle}{M}\right]\right.$$

$$+ w_m^* w_n \sum_j \sum_k w_j w_k^* \left(f_j^r + i f_j^i\right)\left(f_k^r - i f_k^i\right) \exp\left[\frac{-i\pi\langle j - k - m + n, p\rangle}{M}\right]$$

$$\left. + \text{c.c.'s}\right]. \qquad (2\text{-}16)$$

Taking the expected value, the summation over k reduces to those terms

for which $j = k$ since the f_m^r and f_m^i are independent. The first and third terms in Eq. (2-16) are also eliminated because the f_m^r and f_m^i have equal variances. Thus,

$$\sum_p E\left[\frac{\partial I_p}{\partial f_m^r}\frac{\partial I_p}{\partial f_n^r}\right] =$$

$$\sum_p w_m^* w_n \sum_j |w_j|^2 \sigma_f^2 \exp\left[\frac{-i\pi\langle n - m,\ p\rangle}{M}\right] + \text{c.c.} \qquad (2\text{-}17)$$

Finally, the summation over p gives

$$J_{mn}^{rr} = \left[\frac{8c^2\sigma_f^2 M^2 |w_m|^2}{\sigma_N^2}\sum_j |w_j|^2 + \frac{2}{\sigma_f^2}\right]\delta_{mn} \qquad (2\text{-}18)$$

because

$$\sum_p \exp\left[\frac{-i\pi\langle n - m,\ p\rangle}{M}\right] = 4M^2\delta_{mn} \ . \qquad (2\text{-}19)$$

Equation (2-18) is a general expression for one of the submatrices of the Fisher information matrix J given the assumptions above. Similar computations show that $J^{ii} = J^{rr}$ and $J^{ri} = J^{ir} = 0$. In this case, then, J is diagonal and can be analytically inverted to obtain J^{-1}. This is, of course, a result of the discrete Fourier transform nature of Eq. (2-6). Other phase retrieval problems may lead to nondiagonal J matrices which may be difficult or impractical to invert analytically.

Using Eqs. (2-13) and (2-18), the lower bound $(e_{0m})^2$ on the mean-squared error in the estimate of f_m is:

$$e_{om}^2 = \frac{\sigma_f^2}{1 + \dfrac{4c^2\sigma_f^4 M^2 |w_m|^2}{\sigma_N^2} \displaystyle\sum_j |w_j|^2} . \qquad (2\text{-}20)$$

It is, as stated earlier, independent of the phase retrieval algorithm used to estimate f.

The notation of Eq. (2-20) can be simplified by defining a signal-to-noise ratio:

$$SNR = \frac{\{E[cI_p]\}^2}{\sigma_N^2} \qquad (2\text{-}21)$$

where, by Eq. (2-6),

$$E[cI_p] = c\sigma_f^2 \sum_j |w_j|^2. \qquad (2\text{-}22)$$

Equation (2-20) then becomes

$$e_{om}^2 = \frac{\sigma_f^2}{1 + \dfrac{4\ SNR\ M^2\ |w_m|^2}{\displaystyle\sum_j |w_j|^2}} . \qquad (2\text{-}23)$$

As would be expected, the lower bound on the estimate reduces to the a priori variance $(\sigma_f)^2$ if either f_m is not illuminated ($w_m = 0$) or the

SNR is zero. The lower bound also approaches zero as the SNR approaches infinity.

For the case in which the magnitudes of the support constraint w are either zero or one, Eq. (2-20) predicts that, if the support constraint includes a smaller part of the M by M object array (and therefore $\Sigma_j |w_j|^2$ decreases), then the error lower bound increases. This is due to the loss of signal as can be seen from Eq. (2-22). On the other hand, if the SNR (in the measurement plane) is held constant, then Eq. (2-23) predicts that the error bound decreases. This is equivalent to sampling at greater than Nyquist rate in the measurement array in the Fourier domain. The well-known decrease is known as compression gain.

It is known that current iterative phase retrieval algorithms are more successful in converging to a solution for some object support constraints than for others (e.g., for a triangularly-shaped pattern imposed by w, the algorithm more readily finds a solution than for a square pattern) [2.8]. By a solution is meant an object estimate that is as close to agreeing with the measured data and the a priori constraints as possible. In some cases, an algorithm stagnates and produces an output in poor agreement with the data and constraints; such an output should not be considered an object estimate. If there is more than one solution that closely agrees with the data and constraints, the algorithm may find a solution that is different from the true object. There is a tendency for iterative transform algorithms to find solutions more readily for cases guaranteed to have unique solutions (e.g., objects with triangular support constraints). However, when the solution is unique, we have demonstrated that, if a solution is found (i.e., the algorithm does not stagnate in poor agreement with the data and constraints), then the mean-squared error is independent of the shape of the object support constraint [2.9]. From either Eq. (2-20) or

(2-23), it can be seen that, for a given value of $\Sigma_j |w_j|^2$, the lower bound $(e_{0m})^2$ depends only on $|w_m|^2$ and not on the two dimensional distribution of w (the support constraint). The Cramer-Rao lower bound is apparently a measure of the error of algorithms which have found a reasonably good estimate and is insensitive to lack of uniqueness or to algorithm-dependent problems such as stagnation.

2.5 COMPARISON OF LOWER BOUND TO CURRENT ALGORITHM PERFORMANCE

Because it is of interest to compare the Cramer-Rao error lower bound to phase retrieval algorithm performance, computer simulation experiments were performed. Uncorrelated, Gaussian distributed random numbers with zero mean and unit variance were generated and used as the real and imaginary parts of an M by M object array f. This array was multiplied by a binary weighting array w with a triangular shape. The result was zero-filled to array size 2M by 2M, Fourier transformed, and magnitude-squared to give the 2M by 2M Fourier intensity array I. Additional uncorrelated, Gaussian distributed random numbers with zero mean were generated and used as the noise array N. Their variance was varied to achieve different values of the SNR [see Eq. (2-21)]. Without loss of generality, the constant c was set equal to unity. I and N were summed to give the Fourier intensity measurement S.

An iterative Fourier transform algorithm with a combination of the hybrid input-output and error-reduction approaches was used to estimate the object f [2.1, 2.2]. In the object domain, the constraint that no energy fall out outside the binary support constraint w was used. The algorithm was stopped when it had converged to a solution (after about 250 to 1000 iterations). The triangular support was used to minimize any algorithm stagnation problems [2.8]. In general, the complex-valued estimate $\hat{f}$ will differ from the object f by a constant phase. This phase is the angle between the vectors f and $\hat{f}$ and was determined by

taking the dot product between the two vectors. After eliminating this phase difference, the estimation error e, where

$$e^2 = \frac{\sum_m |w_m(\hat{f}_m - f_m)|^2}{\sigma_f^2 \sum_m |w_m|^2} , \qquad (2\text{-}24)$$

was then computed.

One series of computer simulations used M equal to 64 and a right triangular support constraint with base and height equal to 64. The initial estimate of the object input to the iterative algorithm was an array of complex-valued random numbers whose real and imaginary parts were uncorrelated and Gaussian distributed with zero mean and unit variance. Figure 2-2 plots the estimation error e versus the SNR and, for comparison, a corresponding plot of the square root of the normalized Cramer-Rao lower bound, e_o/σ_f [see Eq. (2-23)]. As shown, in a log-log plot, most of the estimation errors lie along a line approximately parallel to the lower bound curve. This behavior holds out to a SNR of 10^{14} at which point the estimation error is about 10^{-7} and the finite precision of the computer is reached.

Another series of simulations used the same support constraint, but the initial estimate was the object to be estimated. Although this starting point cannot be used in any practical application, it is useful in determining the best possible performance obtainable with the algorithm. In this case, estimation error increases in the early iterations and the algorithm converges to a solution consistent with the noisy Fourier intensity measurements. The estimation errors were slightly lower, but nearly identical to those obtained with a random initial estimate. A third series of simulations used M equal to 64 and

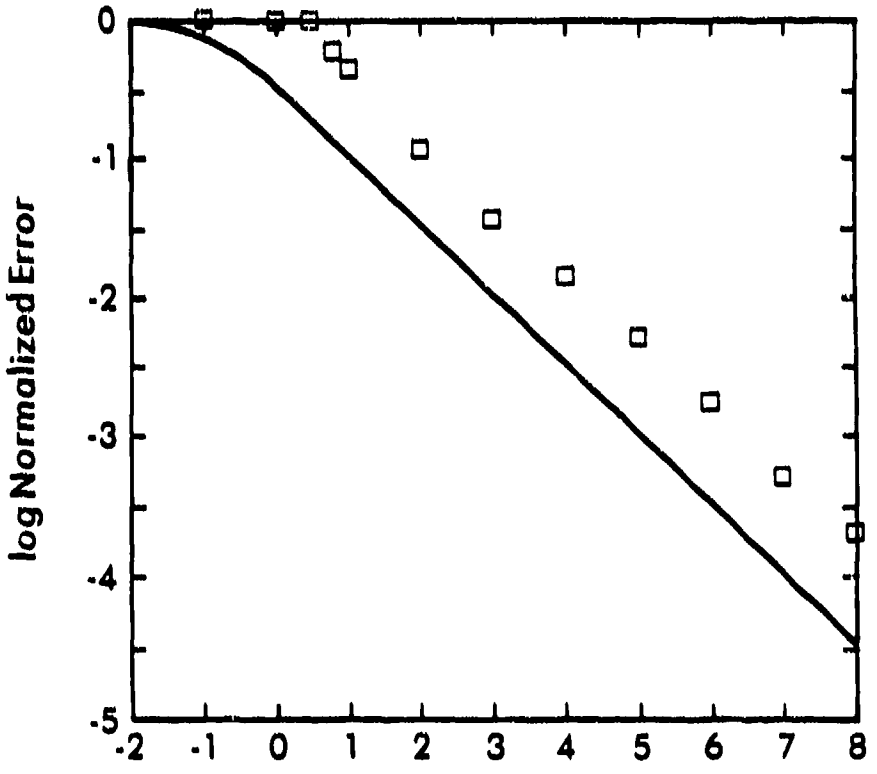


FIGURE 2-2. COMPARISON OF CRAMER-RAO LOWER BOUND (SOLID CURVE) TO SIMULATION ERROR (□'S).

a right triangular support constraint with base and height equal to 16. Again, similar behavior was found, but the separation between the line of estimation error and the lower bound curve increased.

These results encourage the use of estimation theoretic lower bounds in the study of phase retrieval and phase retrieval algorithms. the lower bound did not exceed the simulation errors, thus giving additional confidence in the theoretical developments of Sections 2.3 and 2.4. The slope of the estimation error line is very similar to that of the lower bound curve, thus indicating the lower bound can be used to predict relative performance of the iterative Fourier transform algorithm versus SNR. It is likely both that a greater lower bound can be computed (the Cramer-Rao bound is not necessarily a greatest lower bound) and that an improved algorithm could reduce the estimation error.

2.6 A SECOND PHASE RETRIEVAL PROBLEM

A second phase retrieval problem of interest is one in which noisy complex-valued measurements of the Fourier transform including a random (i.e., degraded) phase are made, rather than real-valued measurements of the Fourier intensity. Using the notational conventions of Section 2.2, the measurement model is

$$S_p = \sqrt{c}\, F_p \exp(i\phi_p) + N_p \qquad (2\text{-}25)$$

where

$$F_p = \sum_m w_m f_m \exp\left[\frac{-i2\pi\langle m,\ p\rangle}{M}\right], \qquad (2\text{-}26)$$

the phase ϕ_p is a random variable, the noise N_p is complex-valued, and, unlike Section 2.2, $p = (p_1, p_2)$; $p_1, p_2 = 0, 1, \ldots, M-1$ (not 2M-1) since the measurements are complex-valued.

The method of Section 2.3 can be applied with the same assumptions as to the statistics of f_m and N_p (with $E[|N_p|^2] = (\sigma_N)^2$) and the relation

$$p(S|f) = \prod_p \int_0^{2\pi} p(s_p|\phi_p, f)p(\phi_p)d\phi_p \tag{2-27}$$

where

$$p(S_p|\phi_p, f) = \frac{1}{\pi\sigma_N^2} \exp\left[\frac{-\left|S_p - \sqrt{c}F_p \exp(i\phi_p)\right|^2}{\sigma_N^2}\right] \tag{2-28}$$

If it is assumed that the phase ϕ_p is uniformly distributed from 0 to 2π (modeling a complete loss of phase information in the measurement of F_p), then (as shown in Appendix B) the lower bound $(e_{0m})^2$ on the mean-squared error in the estimate of f_m is

$$e_{om}^2 = \frac{\sigma_f^2}{1 + \dfrac{(1+a)\ SNR\ M^2\ |w_m|^2}{\sum_j |w_j|^2}} \tag{2-29}$$

where

$$SNR = \left\{ \frac{E\left[\left|\sqrt{c}F_p\right|^2\right]}{\sigma_N^2} \right\}^2 , \tag{2-30}$$

$$a = \frac{2M^2}{\sigma_f^2 \sum_j |w_j|^2} E\left[\frac{|F_p|^2 I_1\left(\frac{2\sqrt{c}|S_p||F_p|}{\sigma_N^2} \right)}{I_0\left(\frac{2\sqrt{c}|S_p||F_p|}{\sigma_N^2} \right)} \right] , \tag{2-31}$$

and I_0 and I_1 are modified Bessel functions of the first kind. Note that Eq. (2-29) is similar in form to Eq. (2-23). Much of the discussion at the end of Section 2.4 is therefore also applicable here. Further investigation of this second phase retrieval problem (such as evaluation of the parameter a) was not undertaken.

2.7 CONCLUSION AND SUGGESTIONS FOR FURTHER RESEARCH

In this investigation of the application of estimation theoretic lower bounds to phase retrieval and image reconstruction problems, the Cramer-Rao lower bound on the mean-squared error in the object estimate from Fourier intensity measurements, given additive noise, Gaussian statistics, and Nyquist sampling, was found. The lower bound approaches the appropriate values in the limits of high and low SNR. Comparison of the lower bound to computer simulations showed that current iterative Fourier transform algorithm performance parallels, but does not achieve the lower bound. The lower bound for the error in estimation from complex-valued Fourier measurements with random phase was also found and is of a similar form to that of the Fourier intensity bound. Further

research should investigate other measurement models (e.g., Fourier magnitude measurements) object domain constraints (e.g., nonnegativity), statistical assumptions (e.g., Poisson noise), and/or other estimation theoretic lower bounds.

REFERENCES

2.1. J.R. Fienup, "Reconstruction and Synthesis Applications of an Iterative Algorithm," Proc. SPIE 373, 147-160 (1981).

2.2. J.R. Fienup, "Phase Retrieval Algorithms: a Comparison," Appl. Opt. 21, 2758-2769 (1982).

2.3. G.B. Feldkamp and J.R. Fienup, "Noise Properties of Images Reconstructed from Fourier Modulus," Proc. SPIE 231, 84-93 (1980).

2.4. H. Van Trees, Detection, Estimation, and Modulation Theory, Part I (Wiley, New York, 1968), pp. 66-73, 79-85, 437-441.

2.5. J.N. Cederquist, S.R. Robinson, D. Kryskowski, J.R. Fienup, and C.C. Wackerman, "Cramer-Rao Lower Bound on Wavefront Sensor Error," Opt. Eng. 25, 586-592 (1986).

2.6. J.W. Goodman, Introduction to Fourier Optics (McGraw-Hill, New York, 1968), pp. 57-62.

2.7. H. Van Trees, "Bounds on the Accuracy Attainable in the Estimation of Continuous Random Processes," IEEE Trans. Inform. Theory IT-12, 298-305 (1986).

2.8. J.R. Fienup, "Phase Retrieval from a Single Intensity Distribution," in Optics in Modern Science and Technology (ICO-13, Sapporo, Japan, 1984), pp. 606-609.

2.9. J.N. Cederquist, S.R. Robinson, D. Kryskowski, J.R. Fienup, and C.C. Wackerman, "Cramer-Rao Lower Bound for Fourier Modulus Wavefront Sensor," Topical Meeting on Signal Recovery and Synthesis II, 86:7 (Optical Society of America, Washington, D.C., 1986), pp. 156-159.

SECTION 3
UNIQUE CLOSED FORM RECONSTRUCTION ALGORITHM

3.1 INTRODUCTION

Since the object's autocorrelation function can be computed from the modulus of its Fourier transform, reconstructing the object from its autocorrelation is equivalent to reconstructing it from the modulus of its Fourier transform. In an earlier effort, it was shown that a unique closed-form algorithm for reconstructing an object from its autocorrelation, which operated in a recursive fashion, was possible for two very special kinds of sampled objects: those fitting within a rectangle with an additional point off one corner of the rectangle and those fitting within a triangle having nonzero corners. This earlier result has been vastly generalized to include sampled objects having supports whose convex hulls have no parallel sides, a very large class of objects. This generalized algorithm, which includes a uniqueness proof, summarized in Appendix C and is described in detail in Appendix D.

Experimental reconstruction results obtained using the algorithm are shown in Section 3.2. Although the present form of the algorithm is very sensitive to noise, limiting its practical use, it has proven to be very valuable in that it suggests useful illumination pattern (support) constraints, as is demonstrated in Section 4.1. Another problem with this reconstruction algorithm is that it explicitly assumes a sampled object, i.e. one consisting of an array of delta functions, and it cannot in its present form be employed for real-world continuous objects. In Section 3.3 an example of how we would experience difficulty in interpreting the uniqueness of the solution for the continuous-object case is shown. One possible way around this problem is to use the quasi-sampling method suggested in Section 3.4.

3.2 EXPERIMENTAL CLOSED-FORM RECONSTRUCTION RESULTS

Autocorrelation data was computer-simulated, including the effects of noise, and images were reconstructed using the closed-form reconstruction algorithm described in Section 3.1 and Appendices C and D.

For each reconstruction experiment an object $f(x,y)$, fitting within a triangular support was Fourier transformed:

$$F(u,v) = \mathscr{F}[f(x,y)] \tag{3-1}$$

The intensity (squared modulus), $|F(u,v)|^2$, of the Fourier transform was computed, and it was scaled in intensity so that the total integrated intensity became equal to a given number of photons,

$$N_p = \sum_{uv} |F(u,v)|^2. \tag{3-2}$$

Then each intensity sample, $|F(u,v)|^2$, was replaced with a random number, $|F_n(u,v)|^2$, drawn from a Poisson distribution with mean and variance equal to $|F(u,v)|^2$. When $|F(u,v)|^2$ is a large number (≥ 32), then a Gaussian approximation to the Poisson distribution is used. This Poisson noise process simulates the effect of photon (shot) noise on the measured Fourier intensity data. The normalized RMS error (NRMSE) of the data is given by

$$E_{|F_n|} = \left[\frac{\sum_{uv} \left(|F_n(u,v)| - |F(u,v)| \right)^2}{\sum_{uv} |F(u,v)|^2} \right]^{1/2}. \tag{3-3}$$

A noisy autocorrelation was computed:

$$r_n(x,y) = \mathscr{F}^{-1}\left[|F_n(u,v)|^2 \right]; \tag{3-4}$$

and an image, $g_n(x,y)$, was reconstructed using the closed-form reconstruction algorithm. The NRMSE of the reconstructed image is given by

$$E_g = \left[\frac{\sum_{xy} |ag_n(x,y) - f(x,y)|^2}{\sum_{xy} |f(x,y)|^2} \right]^{1/2} \tag{3-5}$$

where a is a constant chosen to minimize the error metric, which accounts for the unknown phase constant associated with $f(x,y)$. It can be shown that the value of a that optimizes the image NRMSE is

$$a = \frac{\sum_{xy} f(x,y)g^*(x,y)}{\sum_{xy} |g(x,y)|^2} . \tag{3-6}$$

Examples of images of objects reconstructed from noisy data, for which the object is a uniform triangle, are shown in Figures 3-1 to 3-3 for various sizes of the object. Figure 3-4 plots the image NRMSE versus the data NRMSE for the images shown in Figures 3-1 and 3-2. Several interesting effects are evidenced from these reconstruction examples. First, the closed-form algorithm is very sensitive to noise. A fraction of a percent error in the data results in several percent error in the image. Second, increased data error results in increased image error, but only in an average sense. Occasionally the image error can be greater when the data error is less, because for a given amount of data error the image error that one gets is highly variable. Depending on the particular realization of the noise in the data, the three corner points will have different amounts of error. Small differences in the error of the corner points can yield large differences in the error of the image since the corner points are used over and over again and the error from them propagates and is magnified as the recursive steps build upon one another. This also gives rise to a third effect: the error for the interior points of the reconstructed image is much worse than

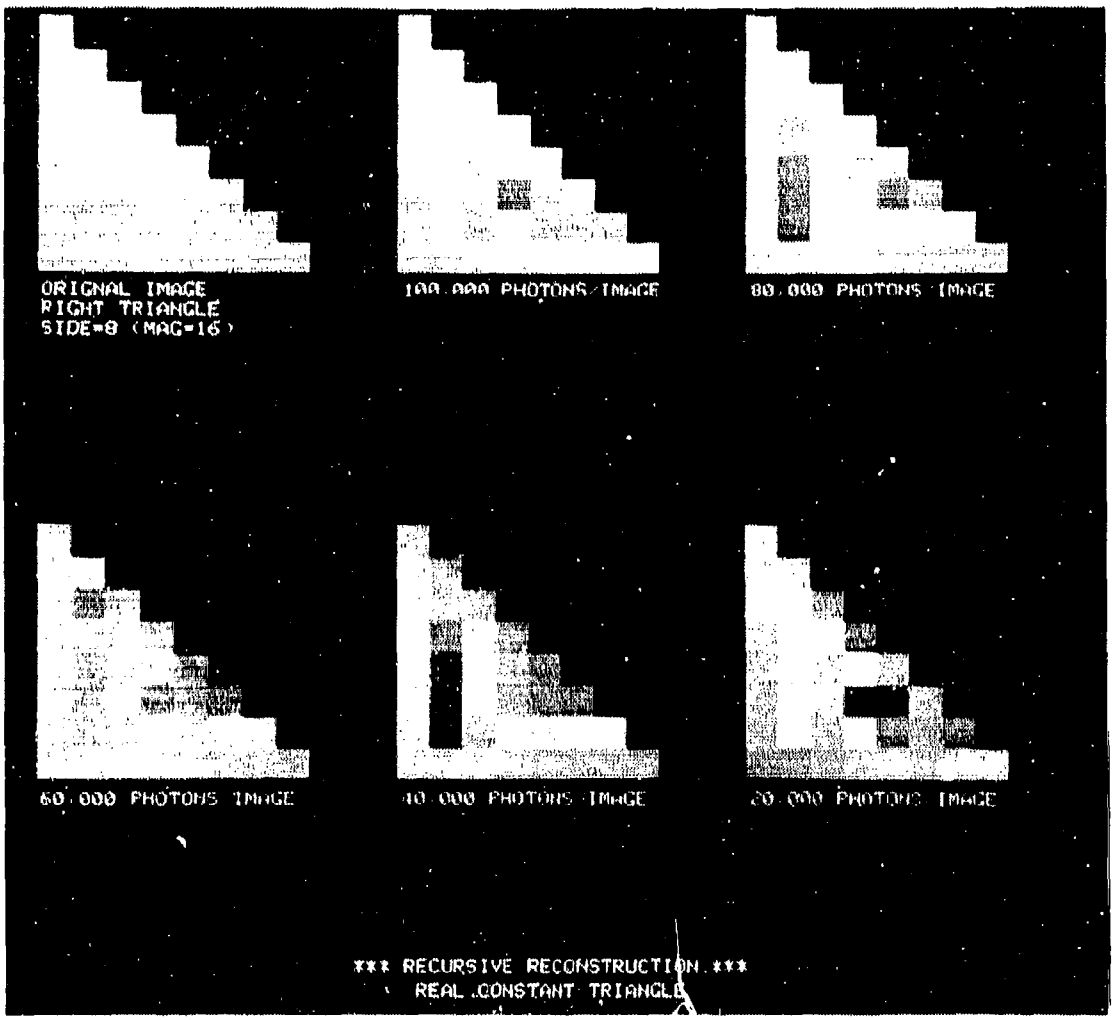


FIGURE 3-1. 8 x 8 OBJECT AND IMAGES RECONSTRUCTED BY THE CLOSED-FORM RECURSIVE ALGORITHM. Number of photons listed in the Fourier intensity domain.

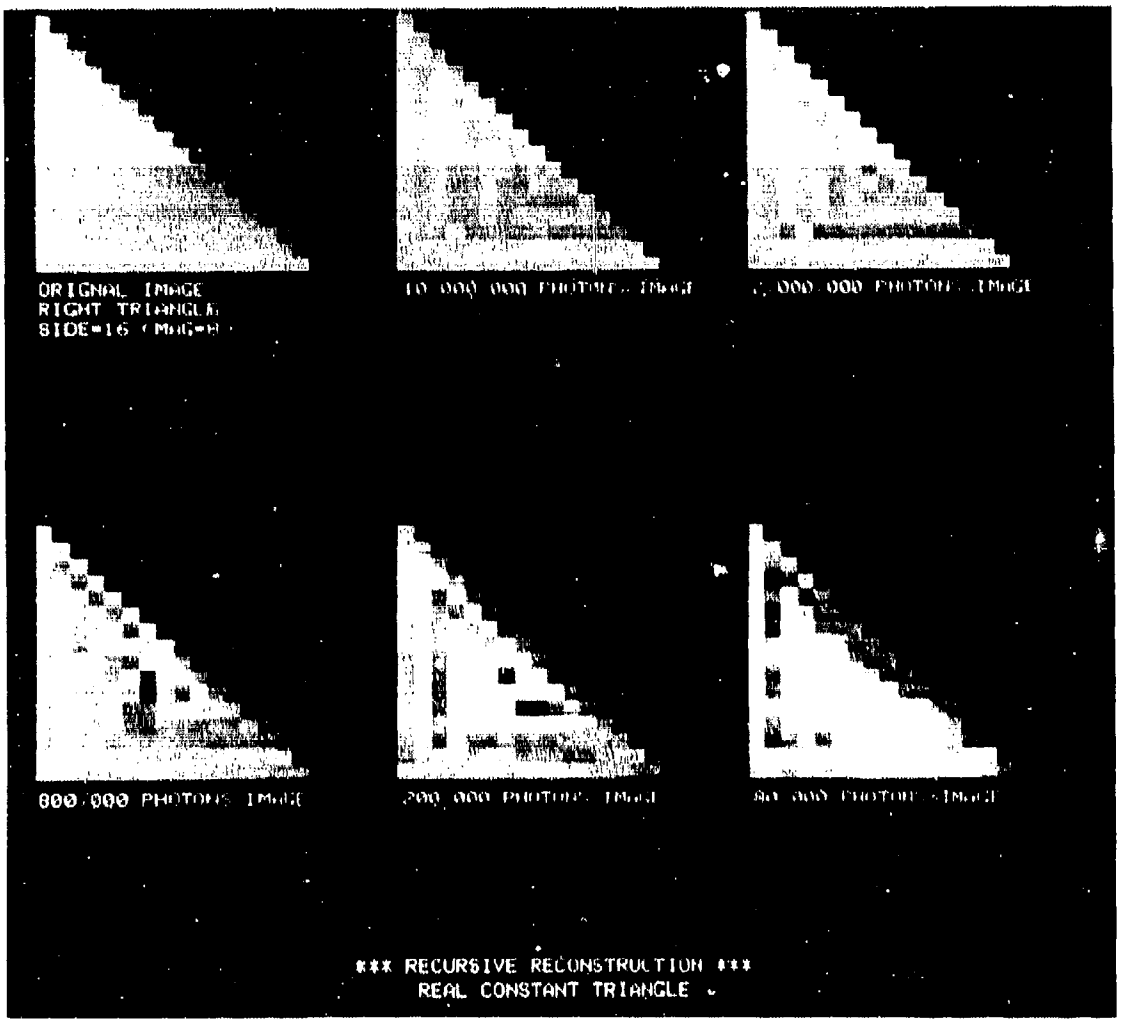


FIGURE 3-2. 16 x 16 TRIANGULAR OBJECT AND IMAGES RECONSTRUCTED BY THE CLOSED-FORM RECURSIVE ALGORITHM.

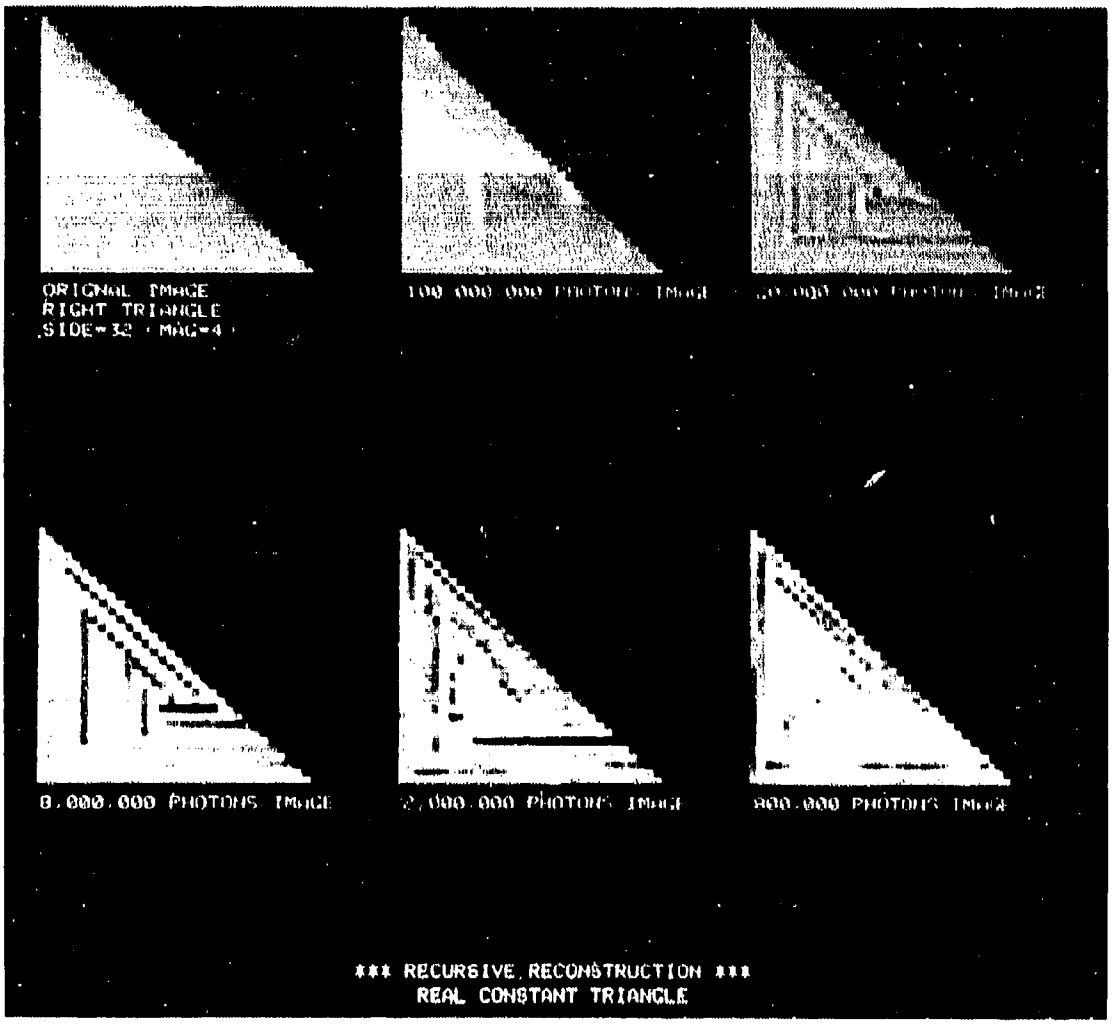


FIGURE 3-3. 32 x 32 TRIANGULAR OBJECT AND IMAGES RECONSTRUCTED BY THE CLOSED-FORM RECURSIVE ALGORITHM.

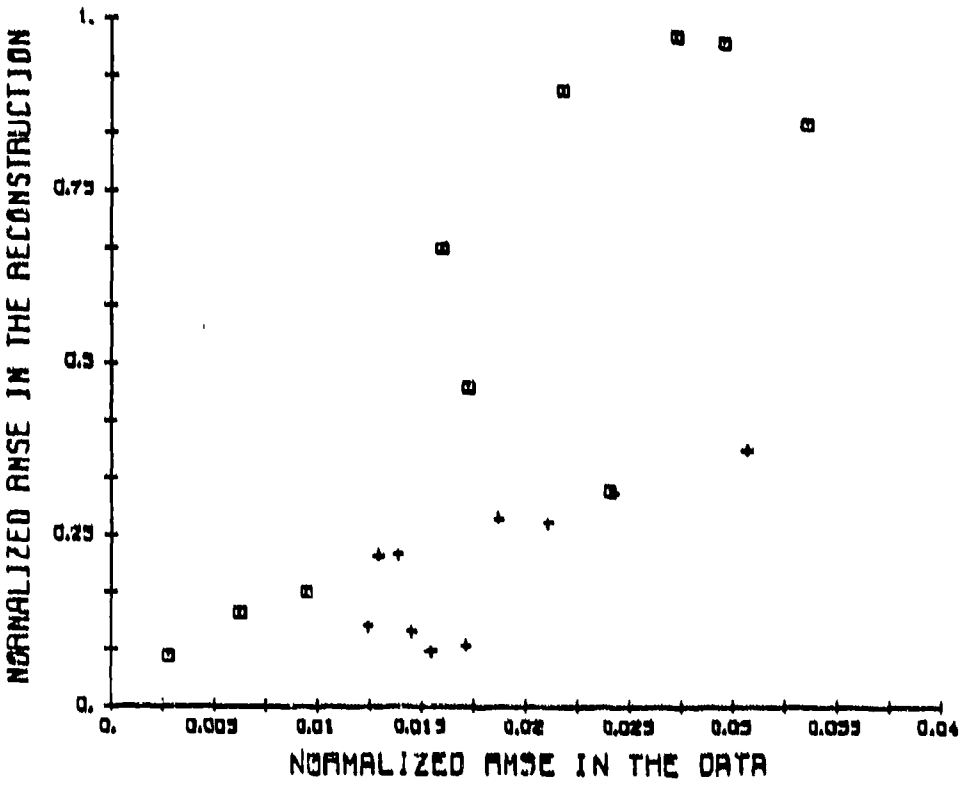


FIGURE 3-4. NRMS ERROR OF THE RECONSTRUCTED IMAGE VERSUS NRMS ERROR OF THE DATA. Crosses are for the 8 x 8 triangles and squares for the 16 x 16 triangles.

the error of the edge points in the image. Fourth, for similar reasons the error of larger images is far worse than the error of smaller images. Fifth, a stripe pattern parallel to each of the edges tends to occur. This happens for the following reason. Suppose that one of the three reconstructed corner points is brighter than it should be. Then the opposing edge of the image, the computation of which involves division by the corner point, will tend to be too dark. Then the next inward row (or column) from the edge, the computation of which involves subtraction of terms involving the edge, will tend to be too bright, etc.

How badly this striping effect affects the interpretability of an image was tested by using a picture of an airplane as the object, imbedded in a 32 by 32 triangle. The results of reconstruction experiments from noisy data using the closed-form reconstruction algorithm for this object are shown in Figures 3-5, 3-6 and 3-7. As seen from Figure 3-5, the image can still be discerned through the partially-obscuring striped pattern. Therefore the intelligibility of the image may be understated by the image NRMSE. Figure 3-6 shows the image NRMSE versus the total number of photons for all the noise values tried for this object, and Figure 3-7 shows the same information, but as a function of data NRMSE. Once again, a fraction of a percent error in the data results in several percent error in the reconstructed image.

Since it was suspected that the corners (vertices) play a pivotal role in determining reconstructed image quality, a series of computer reconstruction experiments were performed with the same object modified to have brighter or dimmer corners. Figure 3-8 shows reconstructed images for the 10^8 photons case for an object with all three corners multiplied by the same factor, for a variety of factors. Figure 3-9 shows a plot of the NRMS error of the images as a function of the corner multiplier. When the corner multiplier is unity (i.e. no modification), the NRMS error of the image ranged between about 0.06 to 0.15 (depending

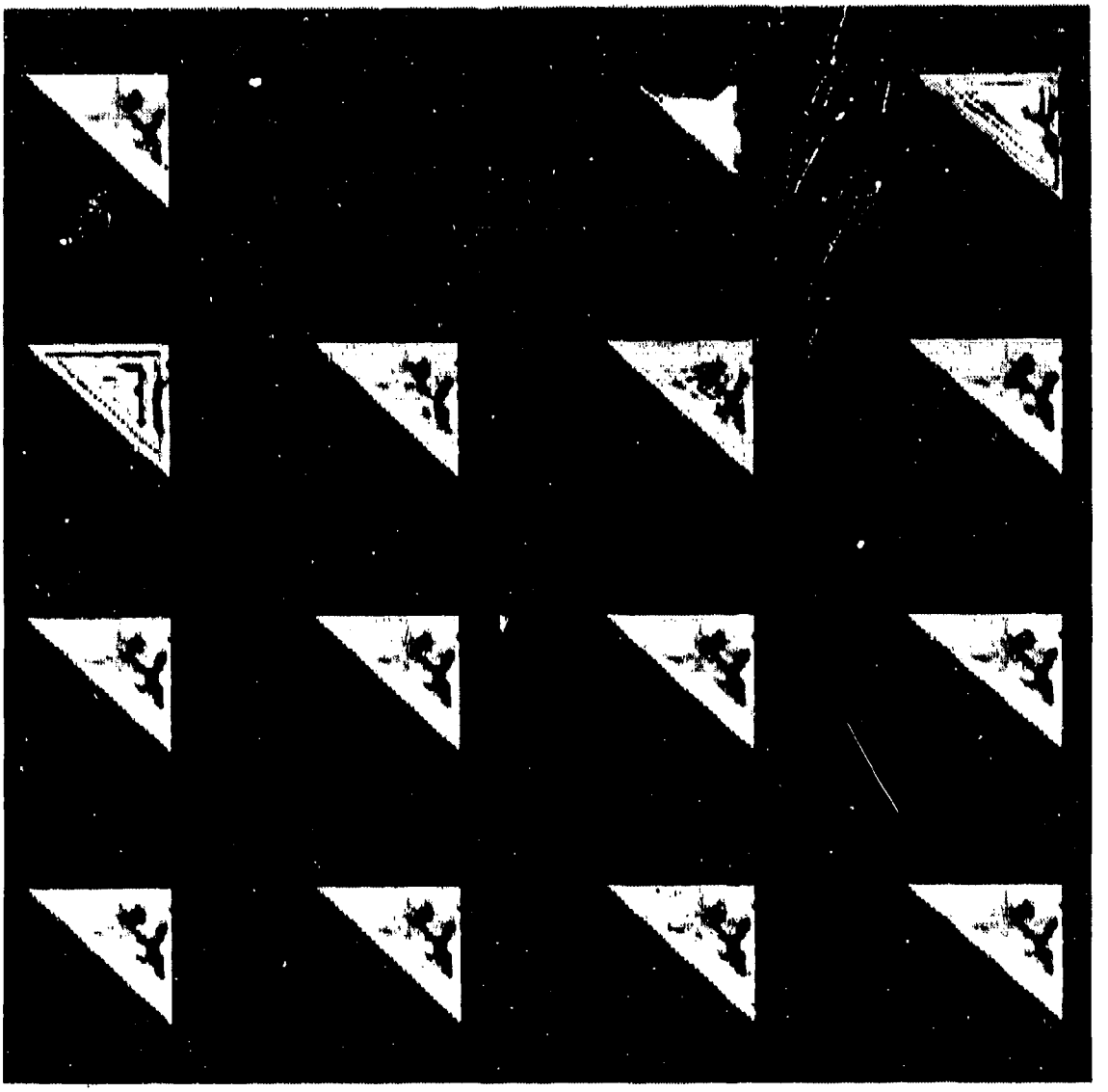


FIGURE 3-5. CLOSED-FORM RECONSTRUCTION OF JET IMAGE, CORNERS UNALTERED. Top left: noise free object; Others: reconstructed images -- number of photons (left to right): Row 1: 10^6, 10^6; Row 2: 10^7, 10^7, 10^8, 10^8; Row 3: 10^9, 10^9, 10^{10}, 10^{10}; Row 4: 10^{11}, 10^{11}, 10^{12}, 10^{12}. Note $10^8 \rightarrow$ 0.0107 NRMSE.

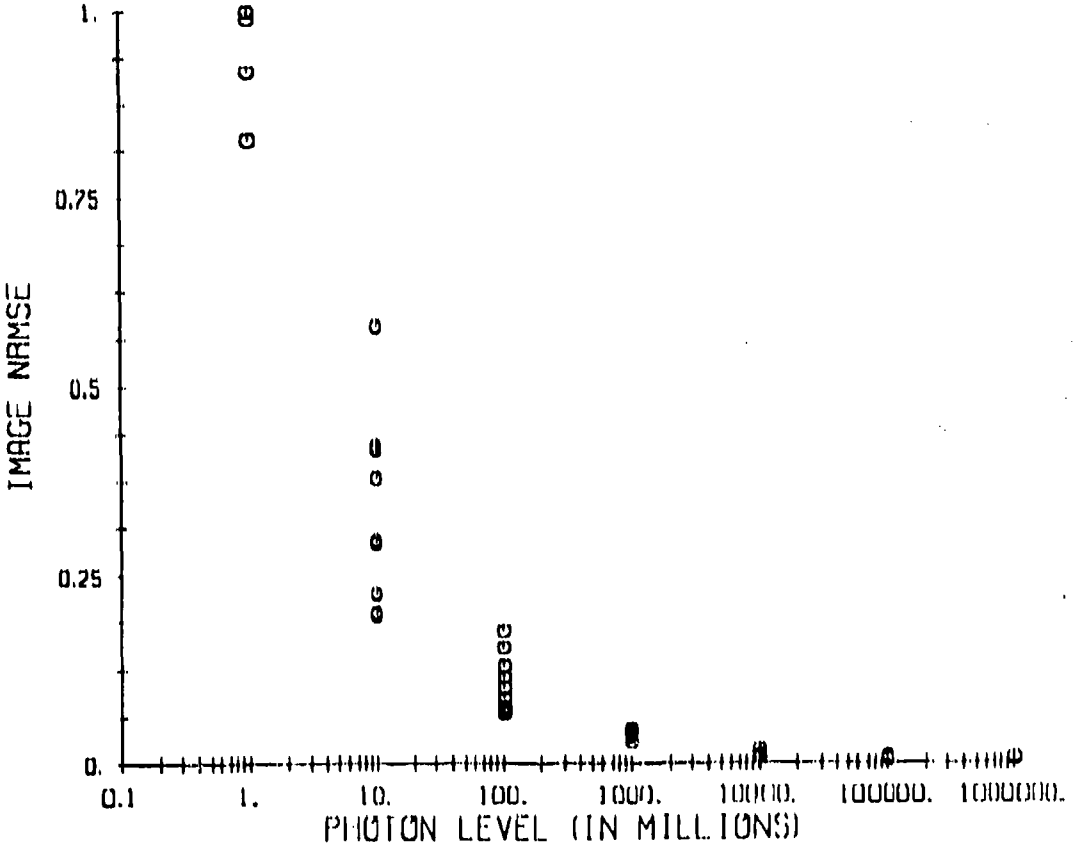


FIGURE 3-6. CLOSED-FORM RECONSTRUCTION ERROR VS LIGHT LEVEL.

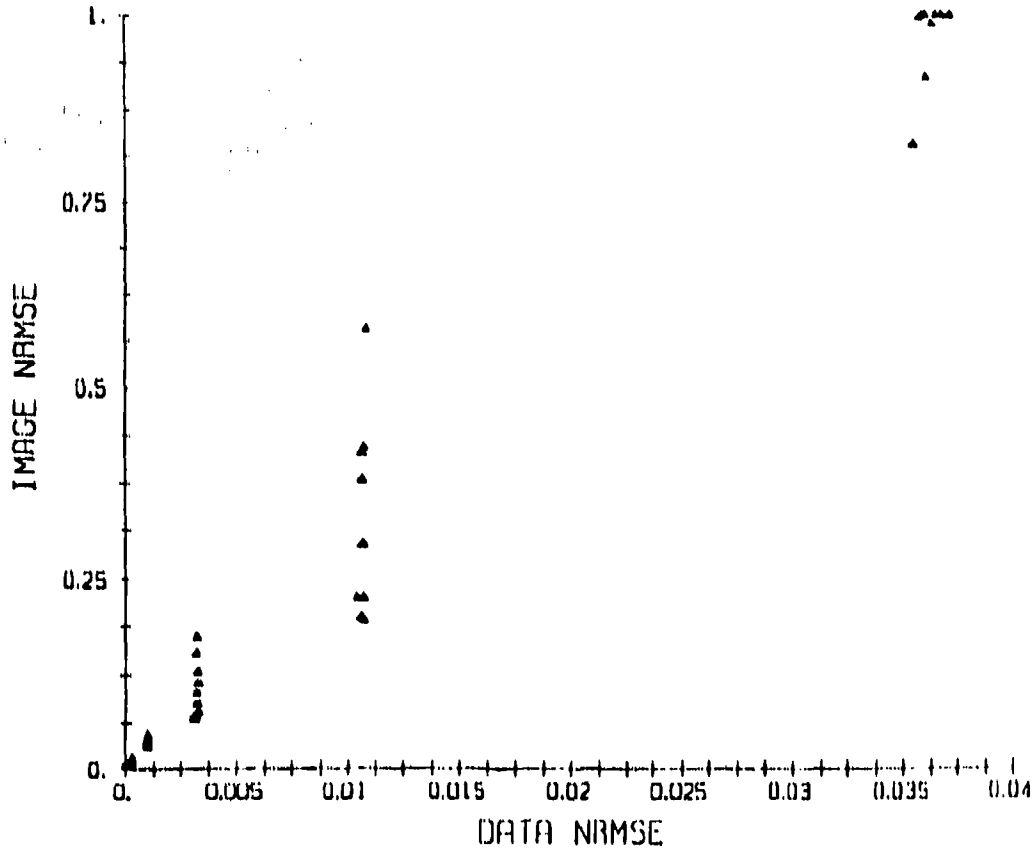


FIGURE 3-7. CLOSED-FORM RECONSTRUCTION ERROR VS DATA ERROR.

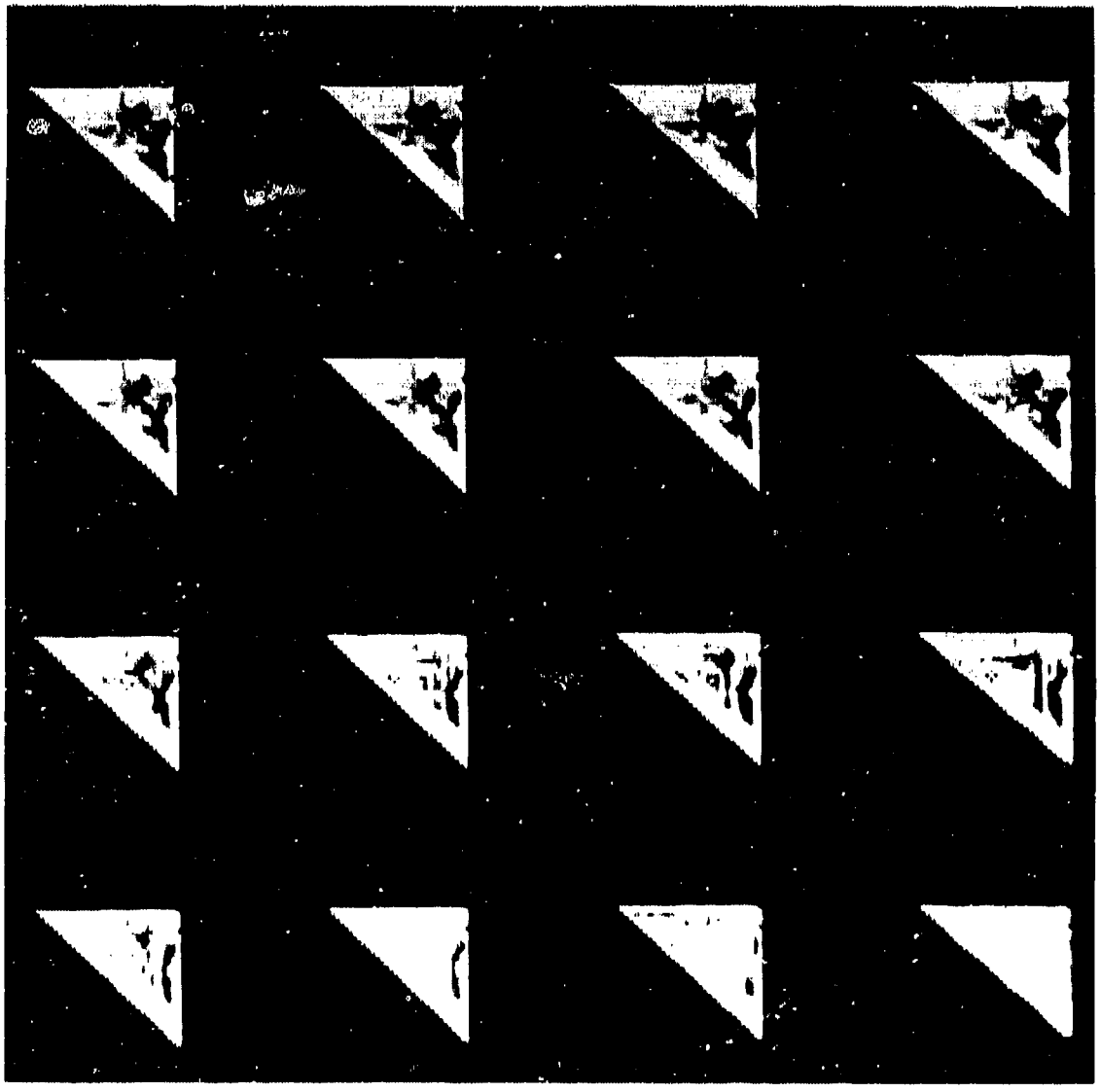


FIGURE 3-8. CLOSED-FORM RECONSTRUCTION CHANGING ALL THREE CORNERS, 10^8 PHOTONS. Reconstructed images -- factor multiplying corners (left to right): Row 1: 64, 32, 16, 8; Row 2: 4, 2, 1, 1; Row 3: 0.9, 0.9, 0.85, 0.8; Row 4: 0.75, 0.71, 0.5, 0.25.

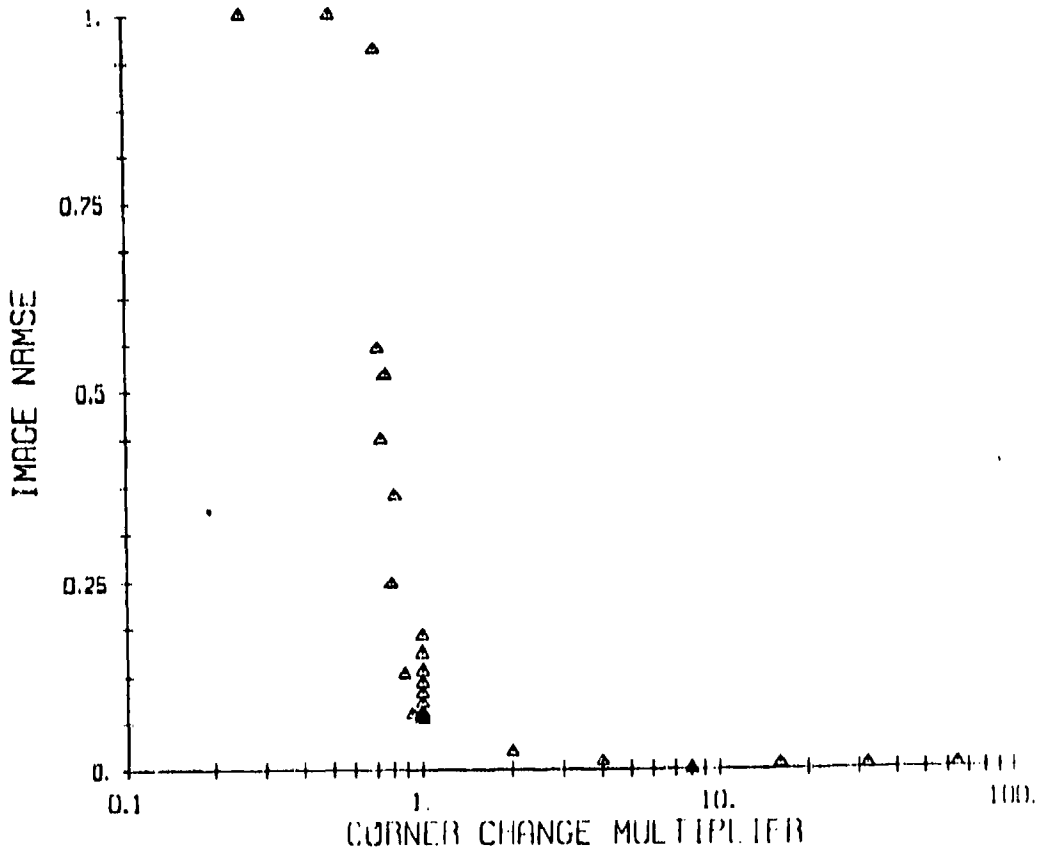


FIGURE 3-9. CLOSED-FORM RECONSTRUCTION ERROR VS FACTOR MULTIPLYING THREE CORNERS, FOR 10^8 PHOTONS.

on the realization of the noise process). For larger values of the corner multiplier the error decreased substantially, and for smaller values it increased dramatically. Similar results are shown in Figs. 3-10 and 3-11 for the case of 10^7 photons (which are noisier than the 10^8 photons case, as expected).

Figures 3-12 through 3-15 show the reconstruction results when only a single corner (the bottom corner) of the object is modified. These results show that when a single corner is brighter, then the portion of the reconstructed image opposite the brighter corner is reconstructed with greater fidelity than the rest of the image. As the brightness of a single corner increases, at first the quality of the entire image improves somewhat, but eventually the quality of the image as a whole degrades when the corner becomes too bright. A possible explanation for this is that as the one corner becomes very bright, the other corners have an increasingly smaller fraction of the total energy, which causes errors in the parts of the reconstructed image opposite them.

From the results above, we can see that when the corners are as bright as the rest of the object, then the closed-form recursive algorithm is highly sensitive to noise; and if the corners are dimmer, then it gets much worse. This is particularly damaging because realistic illumination patterns will have tapered edges and corners, making the corners much dimmer than the rest of the object. Consequently, performance of the closed-form recursive algorithm would be extremely poor if the support constraint is formed by a tapered illumination pattern.

3.3 CONTINUOUS-SPACE TRIANGULAR SUPPORT

The closed-form recursive algorithm includes a proof of uniqueness for triangular objects of known support defined on a rectangular grid of points (i.e. sampled). It does not immediately follow, however, that in the continuous world triangular objects of known support are unique.

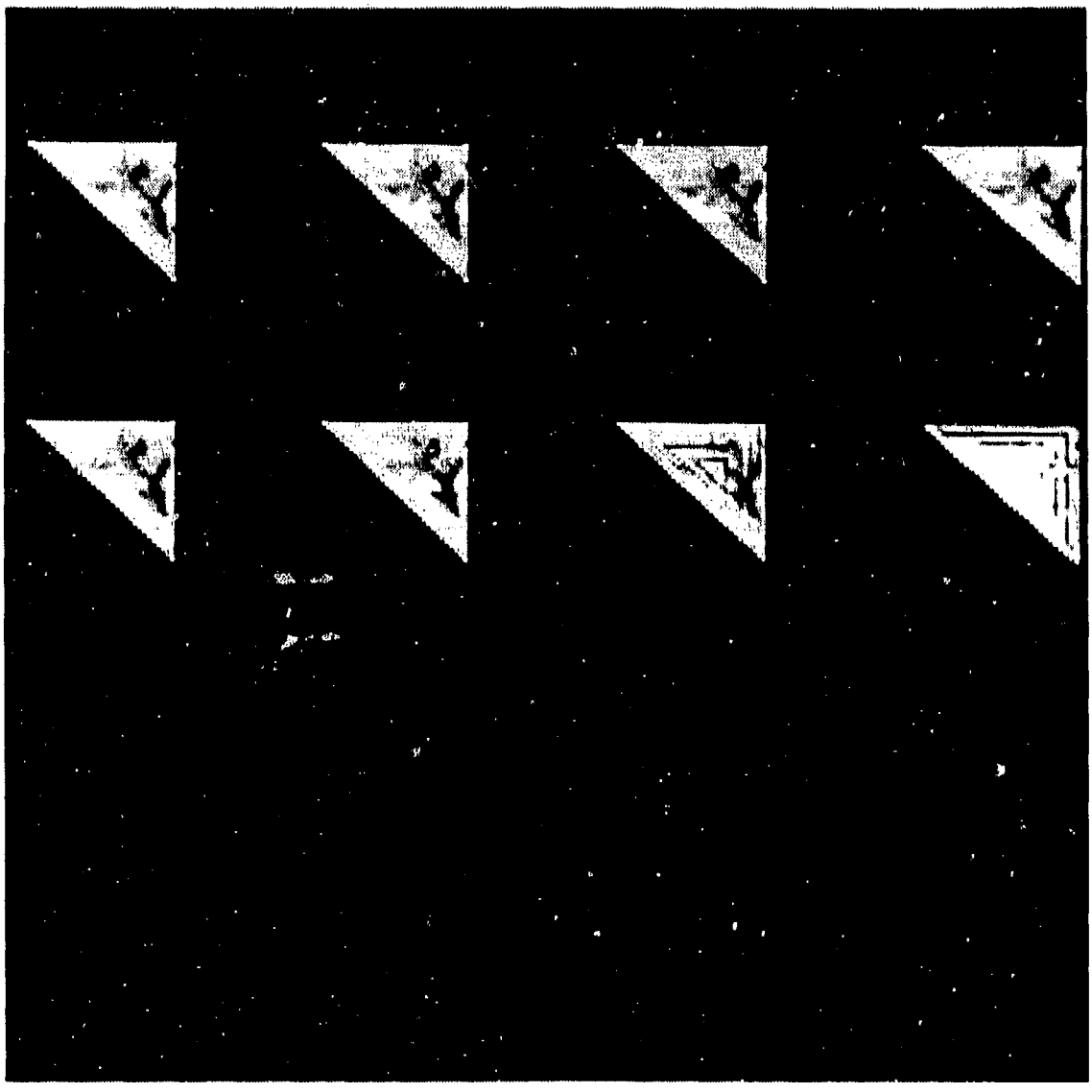


FIGURE 3-10. CLOSED-FORM RECONSTRUCTION CHANGING ALL THREE CORNERS, 10^7 PHOTONS. Reconstructed images -- factor multiplying three corners (left to right): Row 1: 64, 32, 16, 8; Row 2: 4, 2, 1, 0.707.

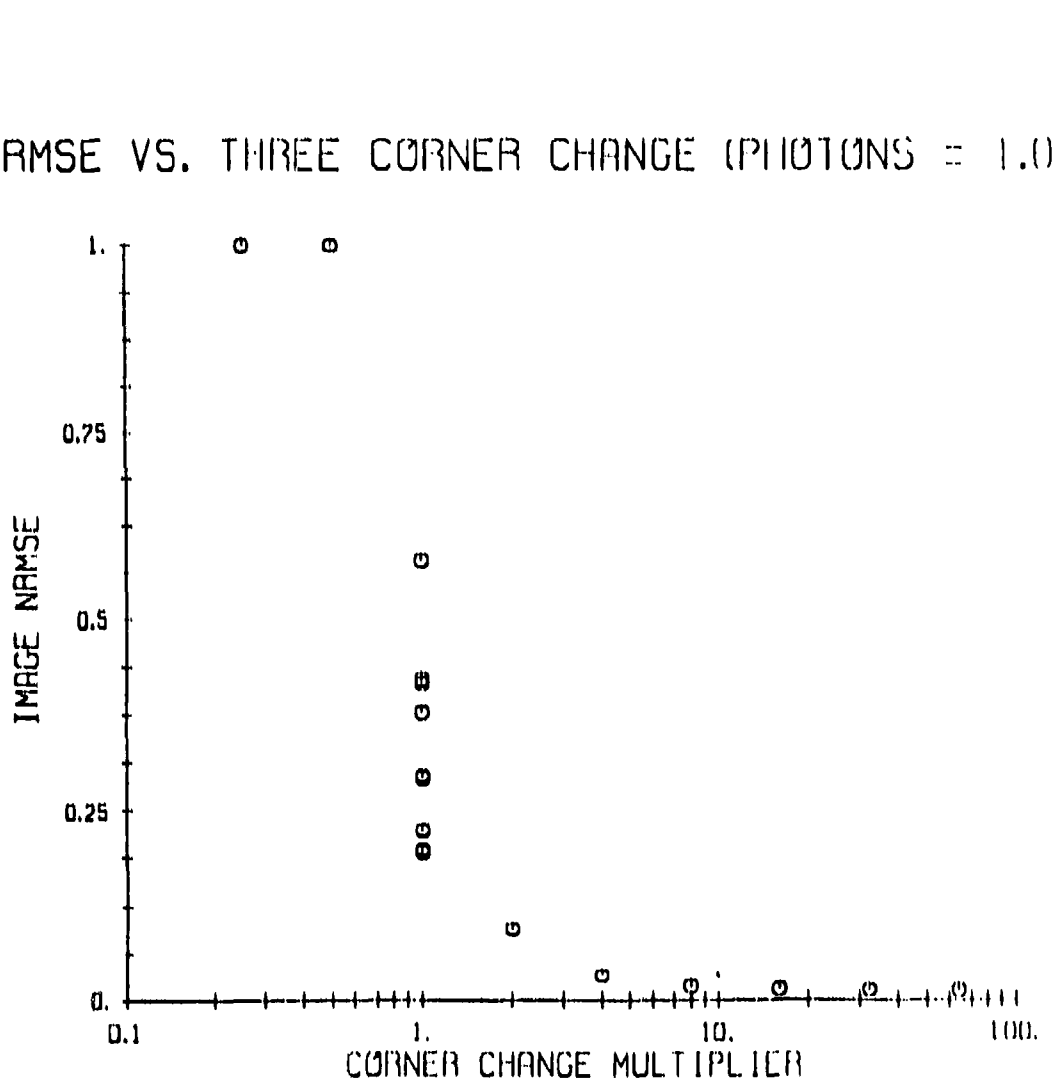


FIGURE 3-11. CLOSED-FORM RECONSTRUCTION ERROR VS FACTOR MULTIPLYING THREE CORNERS, FOR 10^7 PHOTONS.

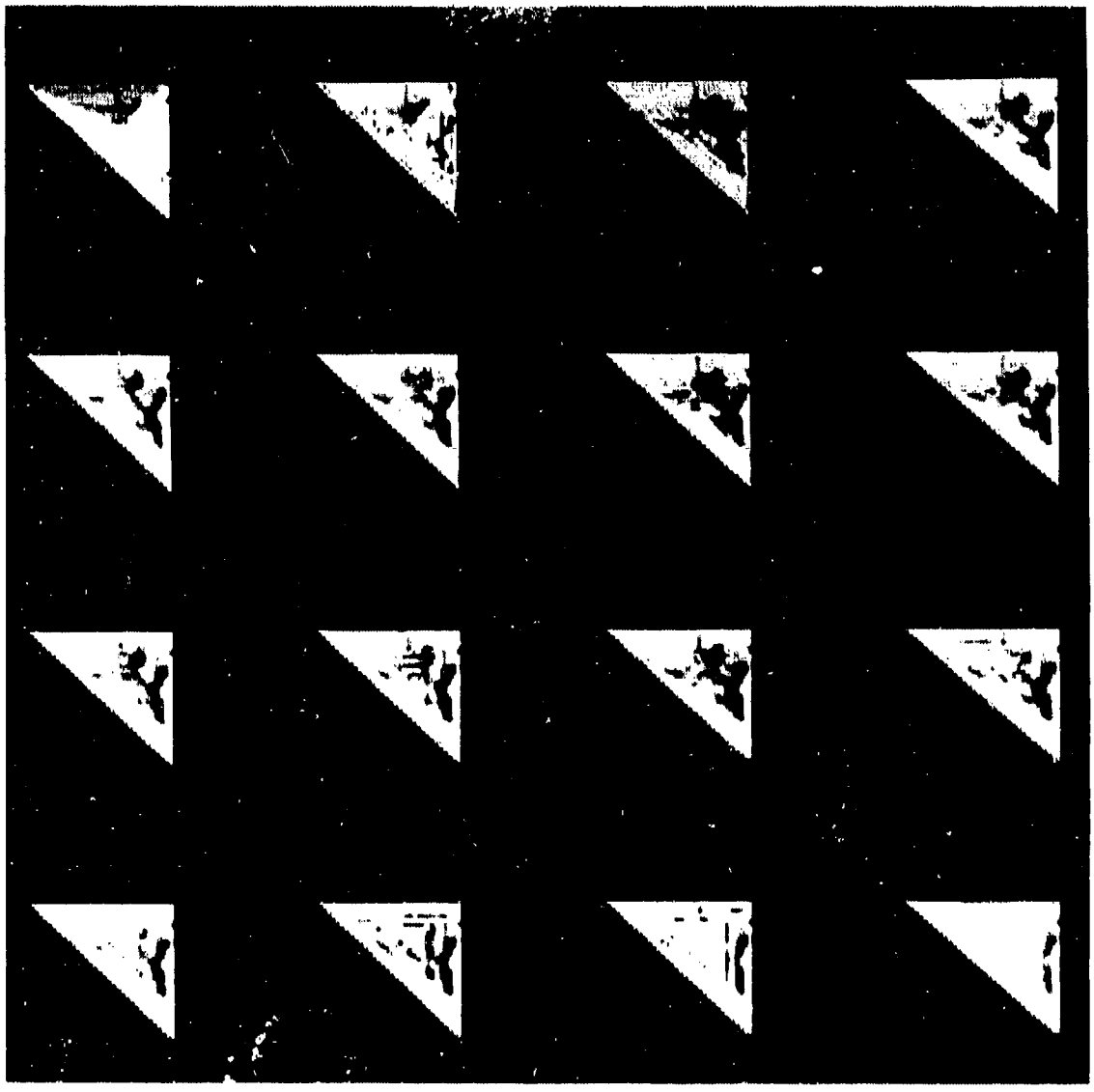


FIGURE 3-12. CLOSED-FORM RECONSTRUCTION CHANGING BOTTOM CORNER, 10^8 PHOTONS. Reconstructed images -- factor multiplying corner (left to right): Row 1: 64, 32, 16, 8; Row 2: 4, 2, 1, 1: Row 3: 0.9, 0.8, 0.71, 0.65; Row 4: 0.65, 0.55, 0.5, 0.25.

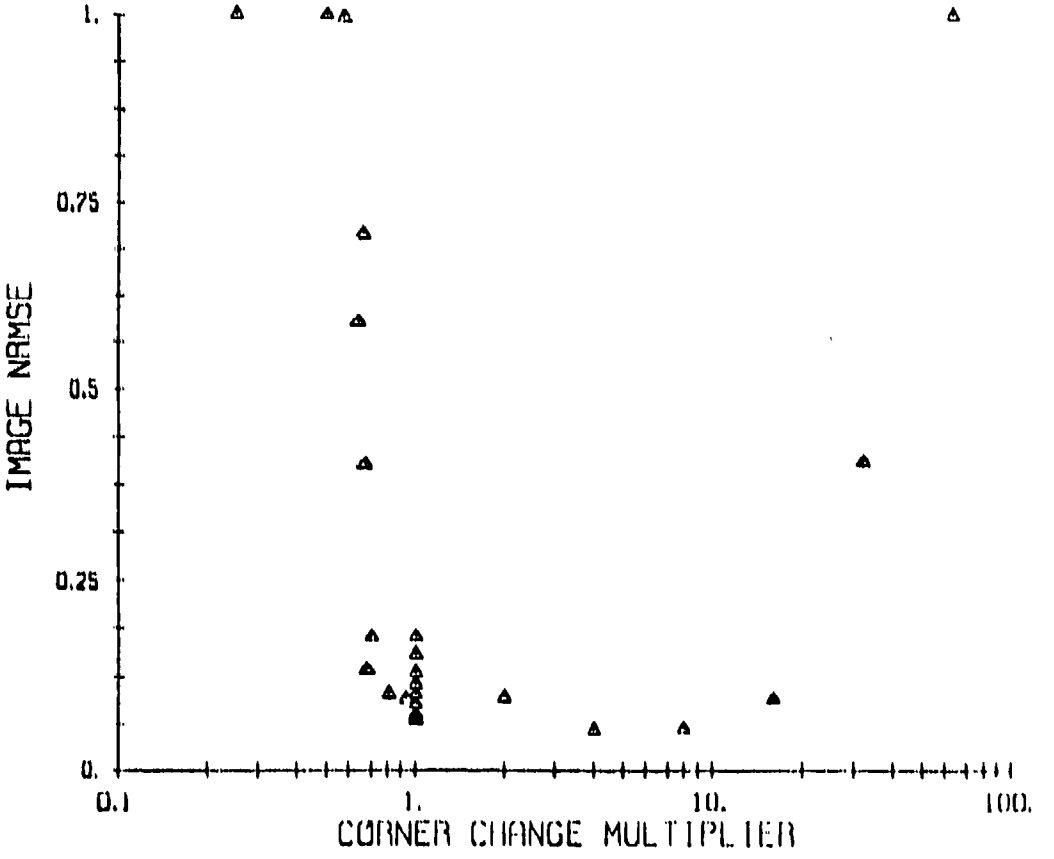


FIGURE 3-13. CLOSED-FORM RECONSTRUCTION ERROR VS FACTOR MULTIPLYING BOTTOM CORNER, FOR 10^8 PHOTONS.

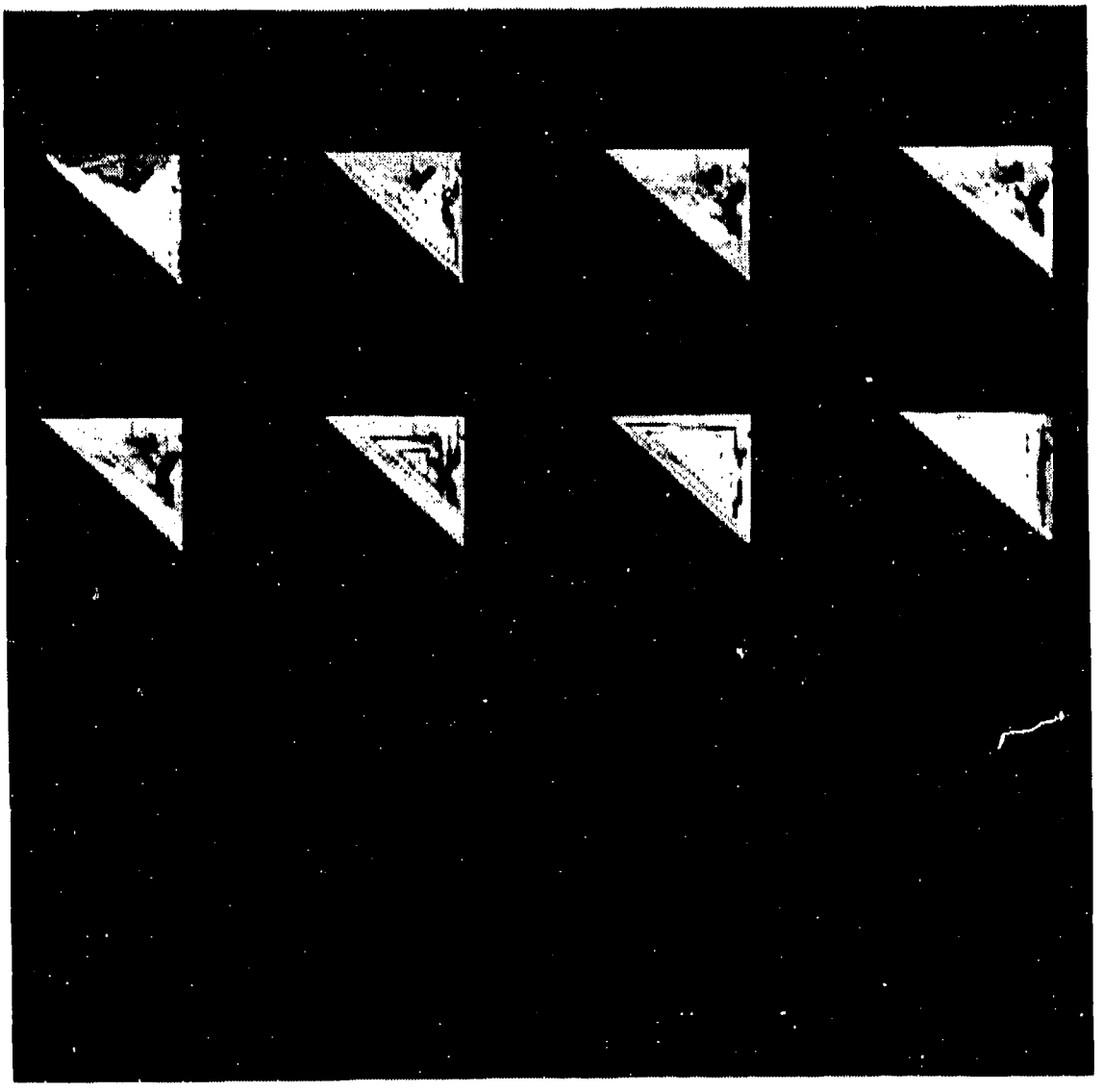


FIGURE 3-14. CLOSED-FORM RECONSTRUCTION CHANGING BOTTOM CORNER, 10^7 PHOTONS. Reconstructed images -- factor multiplying corner (left to right): Row 1: 32, 16, 8, 4; Row 2: 2, 1, 0.71, 0.5.

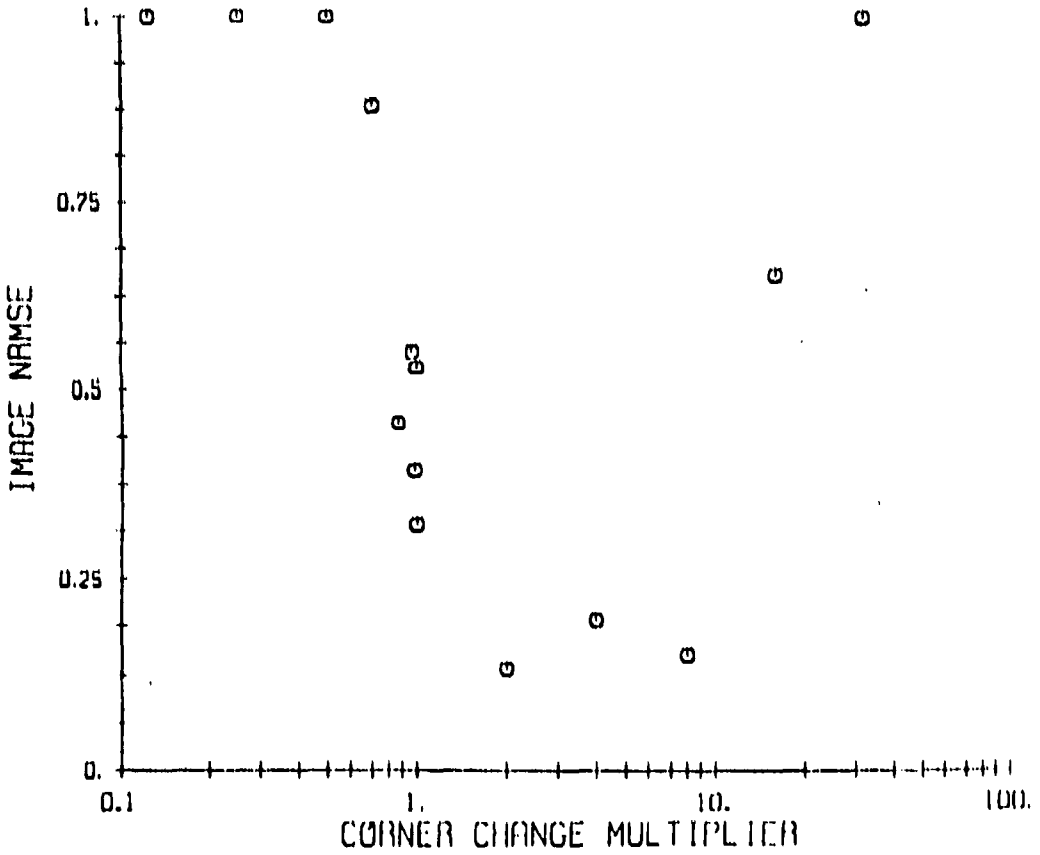


FIGURE 3-15. CLOSED-FORM RECONSTRUCTION ERROR VS FACTOR MULTIPLYING BOTTOM CORNER, FOR 10^7 PHOTONS.

Consider the following two objects:

$$f_1(x,y) = \begin{cases} -x-y, & \text{for } 0 \le y \le 1 \quad \text{and} \quad 0 < x \le 1-y, \\ 2+x-y, & \text{for } 0 \le y \le 1 \quad \text{and} \quad y-1 \le x \le 0, \\ 0, & \text{otherwise} \end{cases}$$

and

$$f_2(x,y) = \begin{cases} 2-x-y, & \text{for } 0 \le y \le 1 \quad \text{and} \quad 0 \le x \le 1-y \\ x-y, & \text{for } 0 \le y \le 1 \quad \text{and} \quad y-1 \le x < 0 \\ 0, & \text{otherwise .} \end{cases}$$

Both objects have the same triangular support

$$T = \{(x,y): 0 \le y \le 1 \text{ and } |x| \le 1-y\}.$$

The objects' Fourier transforms are

$$F_1(u,v) = (1 + iu)\, g(u,v)$$

and

$$F_2(u,v) = (1 - iu)\, g(u,v)$$

where

$$g(u,v) = e^{-iv} \int_0^{-iv} \left[t\, \mathrm{sinc}(ut/2)\right]^2 e^{ivt} dt.$$

Therefore $|F_1(u,v)| = |F_2(u,v)| = (1+u^2)\ |g(u,v)|^2$. Thus we have two different objects with the same support and the same Fourier modulus, contrary to our notion that triangles should be unique. [Although

$f_2(x,y) = f_1(-x,y)$, they are not equivalent solutions with the sense that $f_1(x,y)$ and $f_1^*(-x,-y)$ are].

Note, however, that this is not quite analogous to the discrete triangle case. Since $f_1(x,y)$ approaches +1 as the upper corner (0,1) is approached from the lower left, but approaches -1 as the upper corner is approached from the lower right, we could consider the upper corner to have a value equal to the average of those two limiting values, or zero. But it was assumed for the discrete model that the corners were nonzero. Even though the analogy is not exact, it is close, and this example illustrates the difficulty in translating the discrete-model results to the continuous case, which is the physical case of interest.

3.4 QUASI-SAMPLING ILLUMINATION PATTERN

One problem with the closed-form reconstruction algorithm is its reliance on the object being sampled. In this section, we show that by a special kind of illumination we can approximately achieve the desired sampling effect in the target area.

If one illuminates a target area with four mutually coherent point sources in the far field, at a distance R, given by

$$\delta(u - u_o, v - v_o) + \delta(u - u_o, v + v_o) + \delta(u + u_o, v - v_o)$$

$$+ \delta(u + u_o, v + v_o),$$

we get a field at the target area given by the sum of four plane waves (Fourier transforms of the delta functions):

$$(\lambda R)^{-1} \left\{ \exp\left[\frac{-i2\pi}{\lambda R}(u_o x + v_o y)\right] + \exp\left[\frac{-i2\pi}{\lambda R}(u_o x - v_o y)\right] \right.$$

$$+ \exp\left[\frac{-i2\pi}{\lambda R}(-u_o x + v_o y)\right] + \exp\left[\frac{-i2\pi}{\lambda R}(-u_o x - v_o y)\right]\Big\}$$

$$= (\lambda R)^{-1}\left[\exp\left(\frac{-i2\pi}{\lambda R} u_o x\right) + \exp\left(\frac{-i2\pi}{\lambda R} u_o x\right)\right]\cdot\left[\exp\left(\frac{i2\pi}{\lambda R} v_o y\right) + \exp\left(\frac{-i2\pi}{\lambda R} v_o y\right)\right]$$

$$= 4(\lambda R)^{-1} \cos\left(\frac{2\pi u_o x}{\lambda R}\right) \cos\left(\frac{2\pi v_o y}{\lambda R}\right)$$

which has intensity

$$16(\lambda R)^{-2} \cos^2\left(\frac{2\pi u_o x}{\lambda R}\right) \cos^2\left(\frac{2\pi v_o y}{\lambda R}\right)$$

which has lines of zeros along $x = \lambda R(n + 1/2)/(2u_o)$ and along $y = \lambda R(n+1/2)/(2u_o)$, for $n = 0, \pm 1, \pm 2, \ldots$.

The illumination pattern would be of limited extent which could be modeled by multiplying the above by a slowly varying weighting function defining the field-of-view, $(\lambda R/4)t(x,y)$, so that the entire illumination pattern is

$$w(x,y) = t(x,y) \cos\left(\frac{2\pi u_o x}{\lambda R}\right) \cos\left(\frac{2\pi v_o y}{\lambda R}\right),$$

where there are a number of cycles of the cosines over the extent of $t(x,y)$.

What is accomplished by this is a quasi-sampling of the object. It may be possible to use the closed-form recursive reconstruction algorithm to reconstruct an object, illuminated by $w(x,y)$ above, from

its autocorrelation, or it may be necessary to make modifications to reduce the errors due to approximating this pattern by a true sampling pattern.

Note that by the addition of more plane waves it is possible to get sharper local maxima and broader stripes of low intensity, but at the expense of a more complicated illumination system with phase-stability problems of its own.

In the real world, the four mutually coherent illumination sources may have an unknown relative phasing between them. If the constant relative phases of the four sources are ϕ_1, ϕ_2, ϕ_3 and ϕ_4, then the product of cosines like the equation above occurs only if $\phi_1 - \phi_2 - \phi_3 + \phi_4 = 0$. This implies a stringent stability requirement on the illumination system, but it requires the control of only a single parameter (one piston term) rather than the control of the phase of an entire large aperture.

SECTION 4
CONSTRAINT INVESTIGATION

In this section the various forms of constraints that might be useful for phase retrieval are discussed. Section 4.1 describes results obtained with a variety of support (illumination pattern) constraints. The vast majority of the effort to date concentrated on developing imaging concepts based on the support constraint. Section 4.2 describes other constraints that might also prove useful.

4.1 EFFECT OF ILLUMINATION PATTERN SHAPE

The support, S, of an object is defined as the set of points for which the object is nonzero. For the case of a satellite imaged against the night sky or a ship imaged on calm water with a SAR, the support of the object is basically the filled-in outline of the object. For an airborne or spaceborne sensor looking downward at a general scene, the extent of the object is basically defined by the field of view of the sensor. This latter case may not represent a useful support constraint. However, for an imaging system employing active illumination, the transmitted beam (the illumination beam) can take on a known shape at the plane of the target, and it can be designed to occupy an area smaller than the field of view of the receiver. Then the effective support of the object is the support of the illumination beam pattern. For the case of a SAR, it is assumed that when no phase is available the pulse repetition frequency is at least twice that ordinarily required by Nyquist sampling when phase information is available.

The two most important properties of an illumination pattern are its shape (elliptical, rectangular, polygonal, etc.) and its taper (how slowly it transitions from the bright part of the pattern to where it is effectively zero). As shown in the proposal [4.1, p. 2-29], phase retrieval algorithms are much more effective for some shapes (which we refer to as strong shapes) than for others. Furthermore, phase

retrieval algorithms are more effective for sharp support constraints, i.e. when there is little or no taper to the illumination pattern [4.1, p. 2-25]. Section 5 of this report details the results of our investigation of the effects of tapered illumination patterns and noise and of algorithm improvements that were made for the case of larger amounts of taper. In what immediately follows we discuss the effects of the shape of the support.

In early phase retrieval work there was not an awareness that the support of the object played an important role in the success of phase retrieval. Early successful reconstruction results were for space objects whose supports were naturally non-centrosymmetric [4.2]. Other groups attempted phase retrieval for unnatural objects -- scenes bounded by squares -- and were unsuccessful. Fiddy, Brames and Dainty [4.3] found that the iterative Fourier transform algorithm, although it worked poorly for a rectangular support, worked well for a support consisting of a rectangle plus an extra point just off one corner of the rectangle. This latter support has the special property that any sampled function defined on that support, which is nonzero at the extra point and at one opposite corner, has a Fourier transform that is a nonfactorable polynomial according to Eisenstein's irreducibility theorem. This implies that the phase retrieval problem is unconditionally unique for objects of this type. In retrospect, from those results we can make the crucial connection between three different aspects of the phase retrieval problem: the support of the object, the uniqueness of phase retrieval, and the success of the iterative Fourier transform algorithm.

The trends connecting those three elements, which we have continued to confirm, are the following. First, the support of the object determines whether ambiguities are possible. Second, objects for which uniqueness can be proven are easier to reconstruct by the iterative Fourier transform algorithm than are other objects. The first trend is

amply demonstrated in Appendices C and D which show that sampled objects having known, convex hulls with no parallel sides are unique. The second trend is shown by the reconstruction results [4.1, pp. 2-24 and 2-28] in which objects having known triangular support (which are unique -- see Appendix D) and objects having known supports with separated parts (which even in one dimension are usually unique -- see examples in Section 5) are easily reconstructed while objects with other support constraints, like that of a single ellipse or a single rectangle, are difficult to reconstruct.

The closed-form reconstruction algorithm described in Appendix D may not be practical for use on real-world data, since it requires the objects to be modelled discretely (as a grid of delta functions or sampled points) and it is very sensitive to noise, particularly if the vertex points are dim (see Section 3.2). Nevertheless, it does constitute a uniqueness proof for the types of objects to which it can in theory be applied: objects whose support has a convex hull with no parallel sides. This leads us to consider illumination patterns of this type. Figures 4-1 and 4-2 show examples of reconstruction experiments using illumination-pattern shapes suggested by the uniqueness proof. Figure 4-1(a) shows the modulus of a complex-valued SEASAT SAR image multiplied by a binary pattern (representing the illumination pattern) in the shape of a triangle. Figure 4-1(b) shows the modulus of its Fourier transform (the corresponding signal history) (the Fourier phase was discarded). The iterative Fourier transform algorithm was used to reconstruct an image, the modulus of which is shown in Figure 4-1(c), from the Fourier modulus using the known support pattern. The result is an excellent reconstruction after only 160 iterations. The example of Figure 4-2 is similar, except that a pentagon-shaped illumination pattern and support constraint were used. The result shown in Figure 4-2(d) is after 2990 iterations, when it was still in the process of slowly coverging toward the solution. At this point it strongly resembles the correct object but requires more iterations for complete

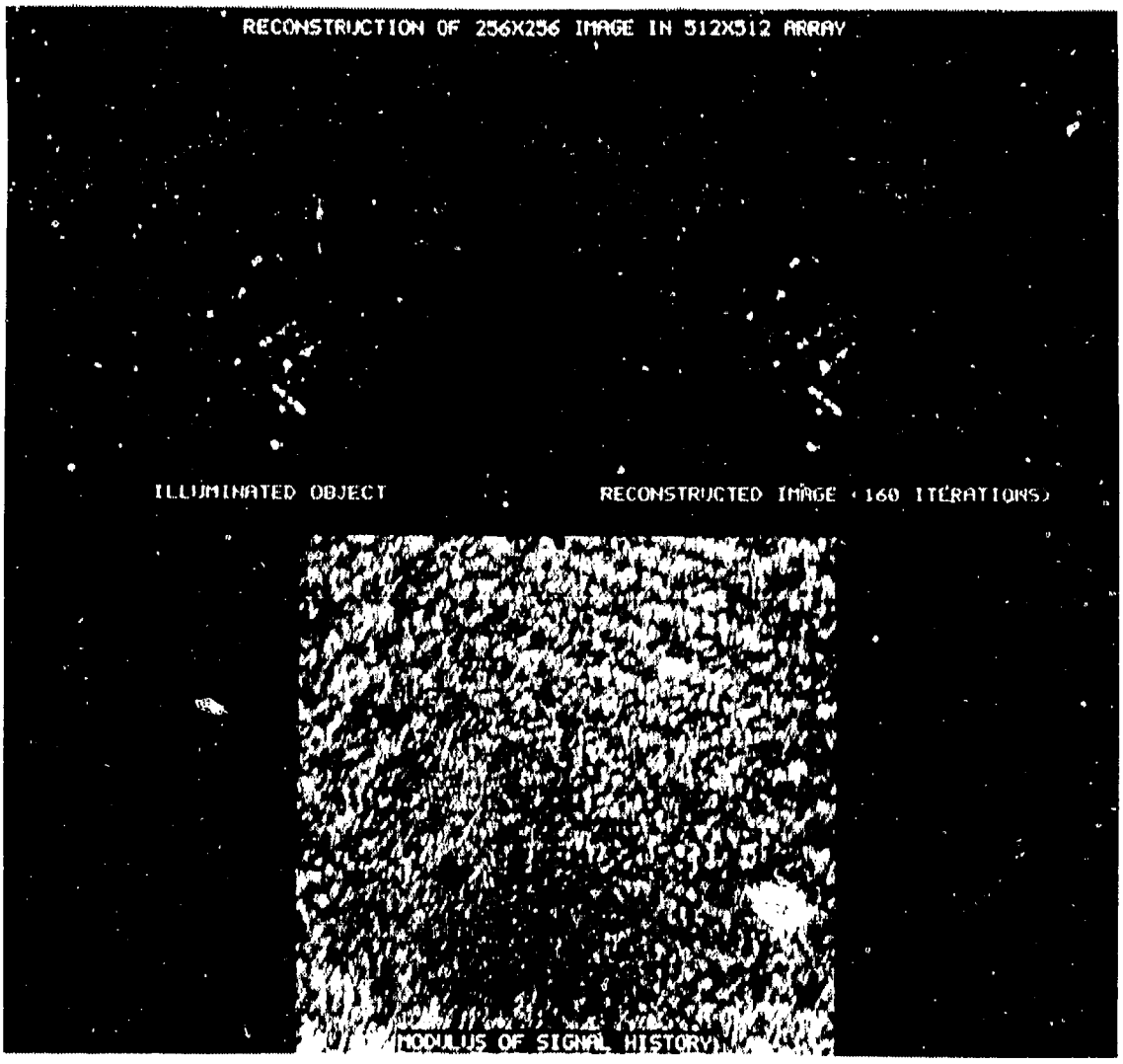


FIGURE 4-1. RECONSTRUCTION EXPERIMENT EMPLOYING TRIANGULAR-SHAPED ILLUMINATION PATTERN. (a) (upper left) modulus of illuminated object, (b) (lower) modulus of signal history, the Fourier transform of the object; (c) (upper right) modulus of image reconstructed from (b) using the iterative transform algorithm with a triangular support constraint.

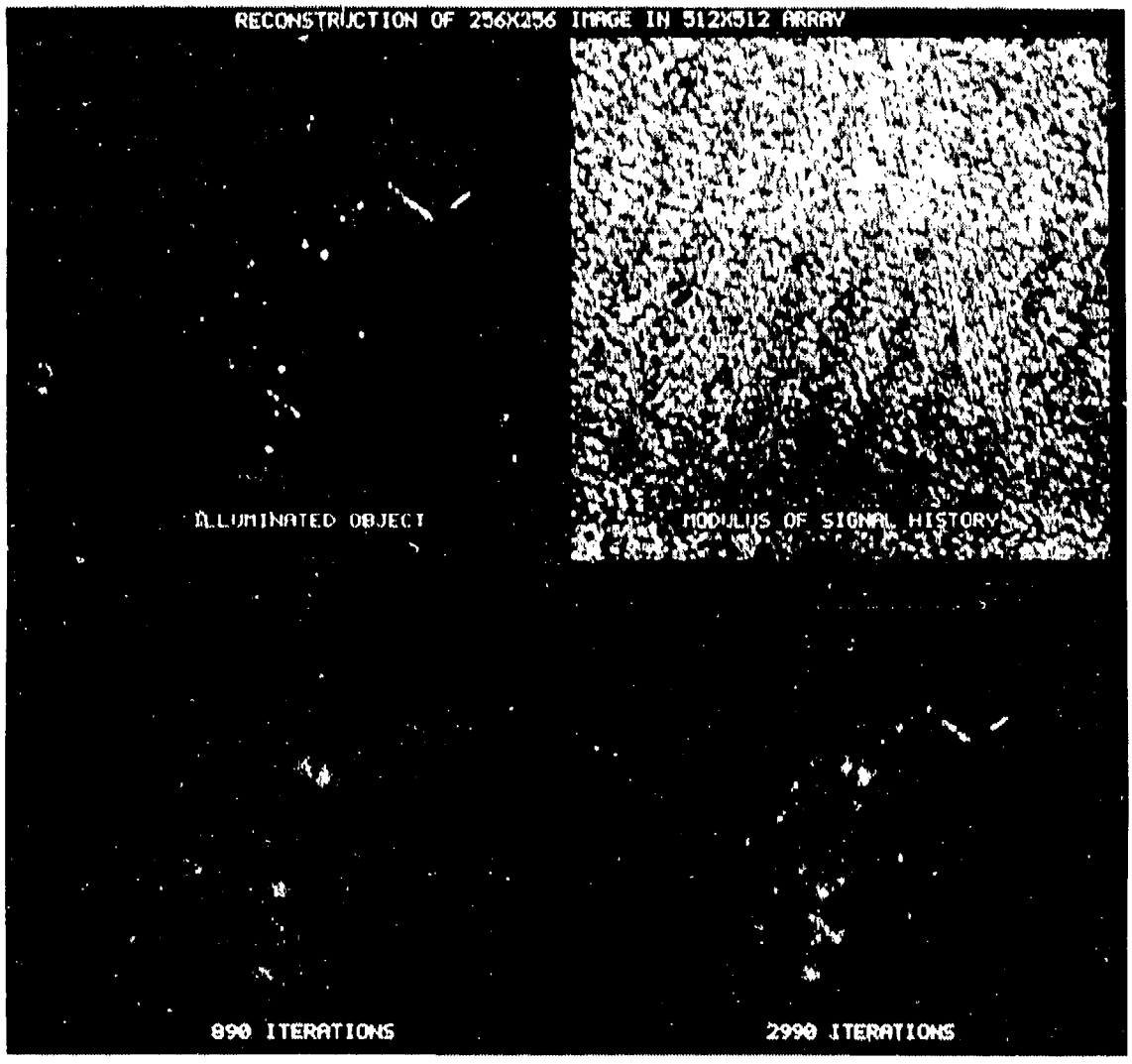


FIGURE 4-2. RECONSTRUCTION EXPERIMENT EMPLOYING PENTAGON-SHAPED ILLUMINATION PATTERN. (a) (upper left) illuminated object; (b) (upper right) modulus of signal history; (c) (lower left) partially reconstructed image after 890 iterations; (d) (lower right) partially reconstructed image after 2990 iterations.

convergence. Given the difficulty in reconstructing complex-valued images with contiguous supports (with the exception of triangular support) [4.1, pp. 2-24 to 2-30], the success of this kind of support constraint would have been difficult to anticipate were it not for the uniqueness proof. Comparison of Figures 4-1 and 4-2 shows that the further the sides are from being parallel, the better. Triangular supports are probably optimal for polygonal supports.

In summary, to date the support constraints found to be most useful for phase retrieval are (a) supports having two or more well-separated parts and (b) supports having convex hulls with no parallel sides. The farther the sides are from being parallel, the better.

4.2 OTHER CONSTRAINTS

Any information that is in a domain other than the domain of the measured data has the potential for being a useful constraint for phase retrieval. Constraints in the domain of the measured data usually just limit the available data and tend not to be useful for phase retrieval. Candidate constraints for the phase retrieval problem are listed in Table 4.1.

Table 4-1
CANDIDATE CONSTRAINTS

Support (illumination pattern)
Nonnegativity
Polarization
Transmitted waveform
Point scatterer in scene
Other scene characteristics

Of these, the support constraint received the most attention, and it is discussed in Section 4.1. The other constraints are described in what follows.

Nonnegativity

The nonnegativity constraint has been very useful in previous phase retrieval efforts [4.2] and also in other image reconstruction problems, such as tomographic reconstruction from incomplete projections and constrained deconvolution. Unlike the support constraint, nonnegativity must exist naturally -- we do not know how to impose it artificially. It naturally occurs in most passive, noncoherent imaging scenarios. The brightness distribution of an incoherently-illuminated reflecting object or a self-emissive object is characterized by a real, nonnegative function (power or photons per unit area). This can be true for actively illuminated objects as well, as long as the illumination is sufficiently incoherent. An exception in which this constraint may not be valid is for passive Doppler imaging as encountered in the PACE program [4.4], for which the aperture function is band-pass. Nonnegativity is not as useful for band-pass systems since the impulse response has very large negative (or complex-valued) sidelobes which convolve the image, destroying its nonnegativity. For the cases in which nonnegativity naturally does occur, it should be relied on heavily as a phase retrieval constraint.

Polarization

Certain kinds of reflecting objects have distinctly different reflectivities for the two different receive polarities (i.e. for same polarity as transmit or for opposite polarity). As an example, corner reflectors reflect either very strongly or very weakly depending on the polarization. Unfortunately, from a single collection it is not immediately obvious how to utilize this information. On the other hand,

if two collections are made simultaneously, one for each polarization, then there is increased possibility of using polarization advantageously. One such possibility would be to use the difference between two degraded images (with measured Fourier phase in the presence of phase errors) to identify point-like reflectors. Then the point-like reflectors could be used in the prominent-point processing described below. Other examples of using polarization may be also be possible.

Transmitted Waveform Type

Early on in the program it was thought that perhaps transmitting a pulse with missing frequency bands might be useful insofar as it would constitute a support-like constraint in the signal history. Upon further examination, it appears that such missing frequencies would primarily result in a loss of data rather than constituting a useful constraint.

A point worth making relating to transmitted wavefront is that the use of phase retrieval techniques may facilitate the use of nonconventional waveforms. As the transmitted waveform departs from the standard set (e.g. the chirp waveform), the availability of hardware that can form the desired waveform in a phase-stable manner may be questionable. By reducing the tolerance on the phase stability of a waveform generator it may be possible to achieve waveforms that would otherwise be very difficult to produce. The reduced tolerance imaging/phase retrieval techniques may provide the means for reducing the phase stability of the waveform generator while maintaining the desired resolution.

Point Scatterers in Scene

Presently, point-like scatterers (prominent points) in the target area are used for correcting small amounts of phase errors in SAR signal histories [4.5, 4.6]. Prominent point processing can also be of great

utility for the case of severe phase errors or when no phase information at all is measured. One particular scenario for phase correction in the presence of large one-dimensional phase errors has already been demonstrated [4.6]. For the case of motion compensation errors in SAR, one has a one-dimensional (azimuth) phase error. This occurs particularly for the case of inverse SAR, for example the radar is ground-based and the (noncooperative) target flies by with a poorly known flight path and rotation. If there exists a dominant prominent point scatterer in a given compressed range cell, then it can be used to calculate the azimuth phase error (taking its phase to be the phase error). The phase errors in all range bins can be corrected by subtracting that phase.

Other Scene Characteristics

The constraints mentioned above are common to large classes of imagery. Also, there may often be additional constraints that exist in specific instances. For example, if the scene has been imaged by another sensor system or by a similar sensor at an earlier time, then these additional images may contain information that can be counted on to appear in the present image and therefore can be used as an a priori constraint. Examples include the known existence of permanent cultural targets or of no-return areas such as lakes or smooth surfaces.

REFERENCES

4.1 "Reduced Tolerance Imaging," ERIM Proposal 653143 to DARPA/TTO, May 1984.

4.2 J.R. Fienup, "Space Object Imaging through the Turbulent Atmosphere," Opt. Eng. 18, 528-534 (1979).

4.3 M.A. Fiddy, B.J. Brames, and J.C. Dainty, "Enforcing Irreducibility for Phase Retrieval in Two Dimensions," Opt. Lett. 8, 96-98 (1983).

4.4 K.K. Ellis, I.J. LaHaie, A.M. Tai, N. Subotic and I. Cindrich, "Passive Interferometric Imaging," Technical Report to AF Wright Aeronautical Labs., Contract No. F33615-84-C-1508 (December 1985).

4.5 J.R. Fienup, "Digital Focusing" in D.E. Klingler et al., "Advanced Synthetic Array Radar Techniques (U)," Third Interim Report, AFAL-TR-77-83, November 1977, pp. 204-233 (SECRET).

4.6 J.C. Dwyer, J.R. Fienup and I. Cindrich, "Hybrid-Optical Prominent Point Processing of Radar Data (U)," 26th Annual Tri-Service Radar Symposium Record, July 1980, p. 285 (SECRET).

SECTION 5
RECONSTRUCTION OF OBJECTS WITH TAPERED ILLUMINATION

5.1 STATEMENT OF PROBLEM

It is well known that knowledge of the support of an object can be a powerful source of information in image-reconstruction problems. By support we mean a compact region outside of which the object is known to be zero, and we denote the set of points that make up the support by the symbol S. In particular, considerable success has been realized in reconstructing an object from its Fourier modulus and a known support [5.1,5.2]. In the reduced-tolerance imaging program an effort is being made to exploit this ability.

Consider an active sensor system that illuminates a target area so that the illumination is confined to a predetermined region. Let $h(x,y)$ be the complex reflectivity of the target:

$$h(x,y) = |h(x,y)|\, e^{i\phi_h(x,y)} . \tag{5-1}$$

Let $w(x,y)$ be the complex illumination fuction:

$$w(x,y) = |w(x,y)|\, e^{i\phi_w(x,y)} . \tag{5-2}$$

We define the effective object as the product of the complex

reflectivity of the target and the illumination function:

$$\begin{aligned} f(x,y) &= w(x,y)\, h(x,y) \\ &= |f(x,y)|\, e^{i\phi(x,y)} . \end{aligned} \tag{5-3}$$

The effective object will now have a support corresponding to the known extent of $w(x,y)$. The intensity pattern of the field emanating from the illuminated target is measured in the far field which may be interpreted

as the squared modulus of the Fourier transform of the effective object. Known phase-retrieval algorithms may then be employed to reconstruct the effective object from the support constraint and the measured Fourier modulus.

Notice that there is some freedom in the choice of the form of the illumination pattern. For example, the shape of the outline of the pattern could be specifically selected to enhance the usefulness of the support constraint. It is known that certain symmetries in object support can create stagnation problems in phase-retrieval algorithms. Consequently the outline of the illumination pattern should have an asymmetric shape. Furthermore, there is some evidence that a support consisting of disjoint regions can be an advantage in phase retrieval. Finally, it is useful to choose an illumination function with a constant modulus over most of the region of illumination thus facilitating the inversion of Eq. (5-3):

$$h(x,y)\Big|_{(x,y)\in S} = \frac{f(x,y)}{w(x,y)} = \left|\frac{f(x,y)}{w(x,y)}\right| e^{i(\phi(x,y)-\phi_w(x,y))} \tag{5-4}$$

when one desires the complex reflectivity of the target without the influence of the illumination pattern.

Unfortunately, the modulus of the illumination pattern will not be binary in practice, but will have some taper associated with it at the edges, due to the effects of diffraction by the aperture of the illuminator. The contrast between an ideal untapered illumination pattern and a more realistic illumination function is illustrated in Figure 5-1. Intuitively we might expect, and experimentally it has been shown [5.1,5.2], that the reconstruction of an object from its Fourier modulus and support would be more challenging for an object with a

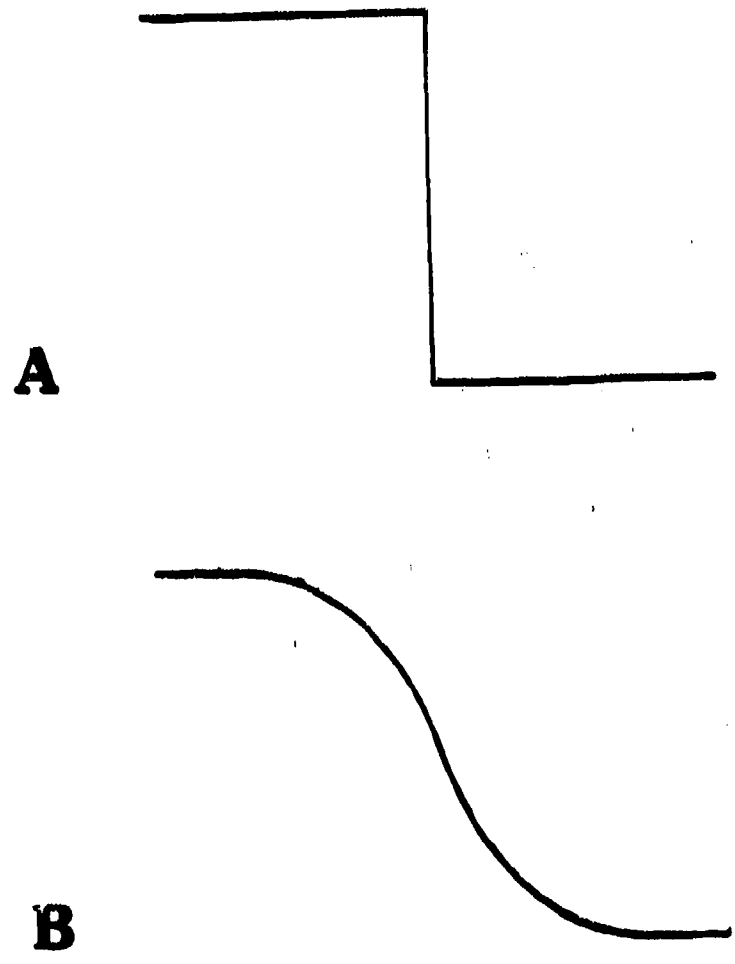


FIGURE 5-1. CROSS SECTIONS OF EDGES OF ILLUMINATION PATTERNS.
A. Ideal binary illumination. B. Tapered illumination.

tapered profile than for one with a sharp profile. It was the purpose of this inquiry to explore this issue via computer simulation and to look for algorithmic modifications that would enhance restoration for this case. For example, it was hoped at the outset that any difficulties incurred by tapered illumination might be offset by the support being disjoint.

5.2 PRELIMINARY SIMULATIONS

We began by exploring the effect of tapered illumination on phase retrieval by means of computer simulation. A pair of disjoint ellipses was used as the basic shape for the illumination pattern. The untapered illumination pattern was assigned a value of unity within the ellipses and zero outside. Taper was introduced by convolving the binary ellipses with a convolution kernel. The normalized kernels used in these preliminary simulations are shown in Figure 5-2. Cross sections of the edge of the resulting illumination patterns are given in Figure 5-3.

As mentioned earlier, it was speculated that disconnected support might help to overcome any problems associated with tapered illumination. For this reason the total illumination pattern was chosen to be two disjoint ellipses. Simulations were performed for objects with differing amounts of illumination taper and differing amounts of separation between ellipses in the illumination pattern. A given simulation was performed by first multiplying complex SEASAT SAR imagery by the given illumination pattern to create an effective object. Because it is the effective object that we try to recover through phase-retrieval techniques, we will henceforth refer to this as the true object. This object was Fourier transformed with an FFT and the Fourier magnitude was retained. The known region of support in the object domain was supplied by hard limiting (thresholding) the illumination pattern with a very small threshold value. A standardized sequence of error-reduction and hybrid input-output iterations [5.3] were then

A

1/16	1/16	1/16
1/16	1/2	1/16
1/16	1/16	1/16

B

1/9	1/9	1/9
1/9	1/9	1/9
1/9	1/9	1/9

FIGURE 5-2. DISCRETE CONVOLUTION KERNELS USED TO ADD TAPER TO BINARY ILLUMINATION PATTERN. A. Center-weighted kernel yields taper #1. B. Evenly-weighted kernel yields taper #2.

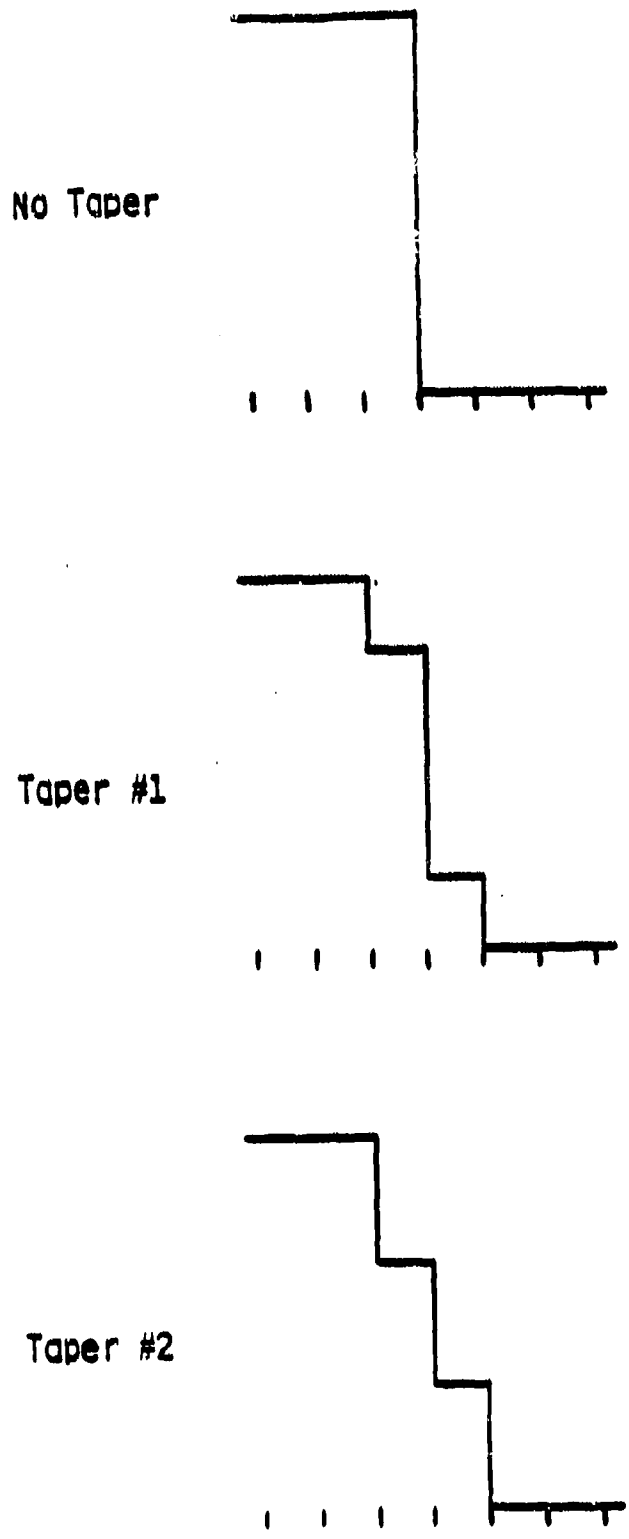


FIGURE 5-3. CROSS SECTIONS OF ILLUMINATION OF TAPER USED IN PRELIMINARY SIMULATIONS

performed to reconstruct the object from its Fourier modulus and support. Convergence for all simulations was monitored by calculating a Fourier domain normalized error metric,

$$E_F^2 = \frac{\sum_{u,v} \left(|G(u,v)| - |F(u,v)| \right)^2}{\sum_{u,v} |F(u,v)|^2} \qquad (5\text{-}5)$$

where $F(u,v)$ is the discrete Fourier transform of the true object $f(x,y)$, and $G(u,v)$ is the Fourier transform of the image estimate. The convergence is portrayed in Figure 5-4 for six kinds of illumination--three amounts of taper, each with two amounts of separation between ellipses. It is important to note that Figure 5-4 is a log-log plot and therefore the behavior of the algorithm becomes horizontally compressed with increasing number of iterations. Figures 5-5, 5-6, and 5-7 give the final reconstructions for each of the cases tested. These results confirm our expectation that increased amounts of illumination taper make the reconstruction process more difficult. In fact, for the case with the largest amount of taper the algorithm convergence appears to have stagnated. This is in spite of the fact that the amount of taper is extremely mild. There are 51 pixels along the major axis of the large ellipse and only two pixels of taper at the edge. Thus convergence of the traditional algorithm appears to be relatively sensitive to illumination taper. It is important to realize that the convergence curves shown in Figure 5-4 correspond to a specific initial estimate and that the convergence behavior could vary when alternative initial estimates are used.

5.3 THE SHRUNKEN-MASK ALGORITHM

In order to explore the reasons for stagnation we created a difference image between the modulus of the true object and that of the restored object for the case of intermediate taper (taper #1). This difference image is bipolar and a bias was added for display in Figure

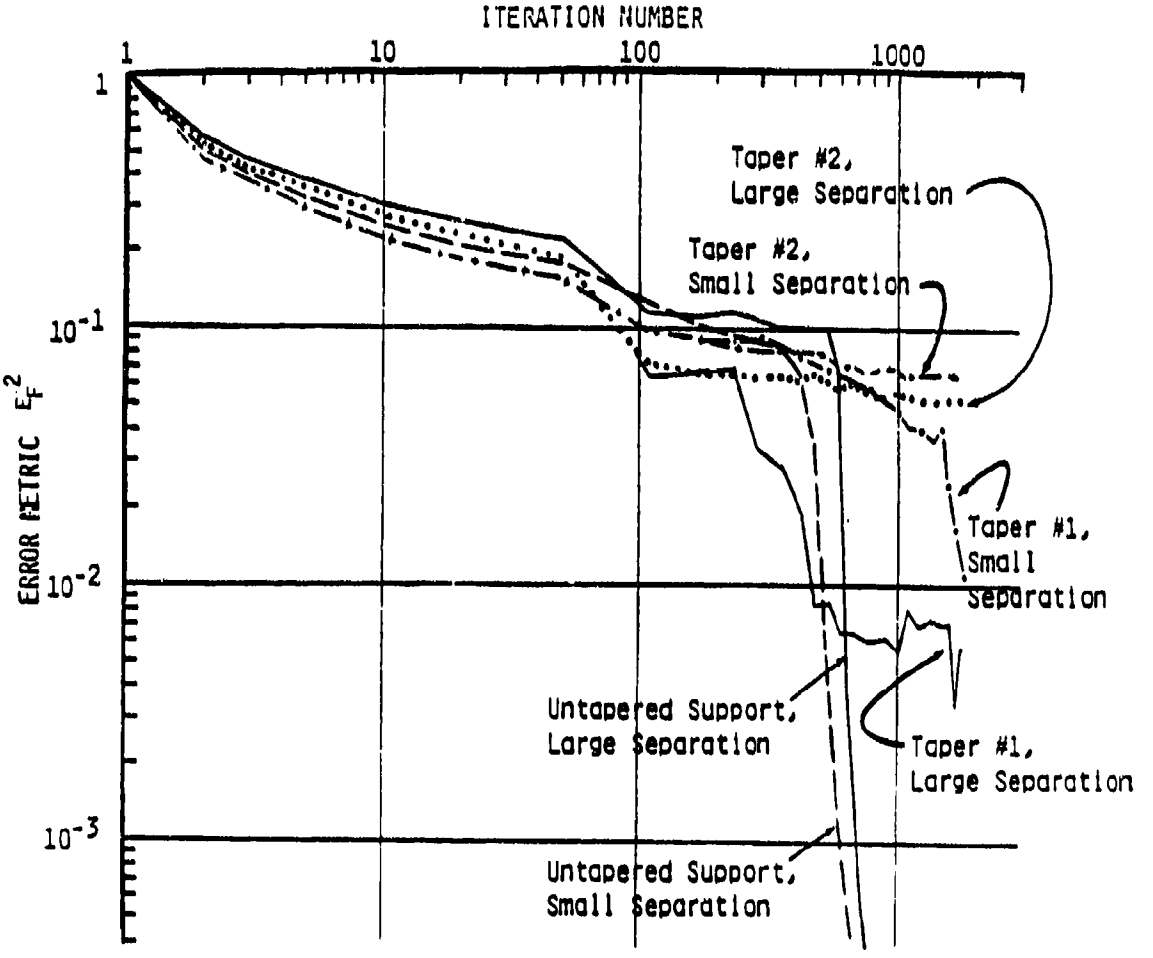


FIGURE 5-4. CONVERGENCE BEHAVIOR AS A FUNCTION OF ILLUMINATION TAPER AND SUPPORT SEPARATION

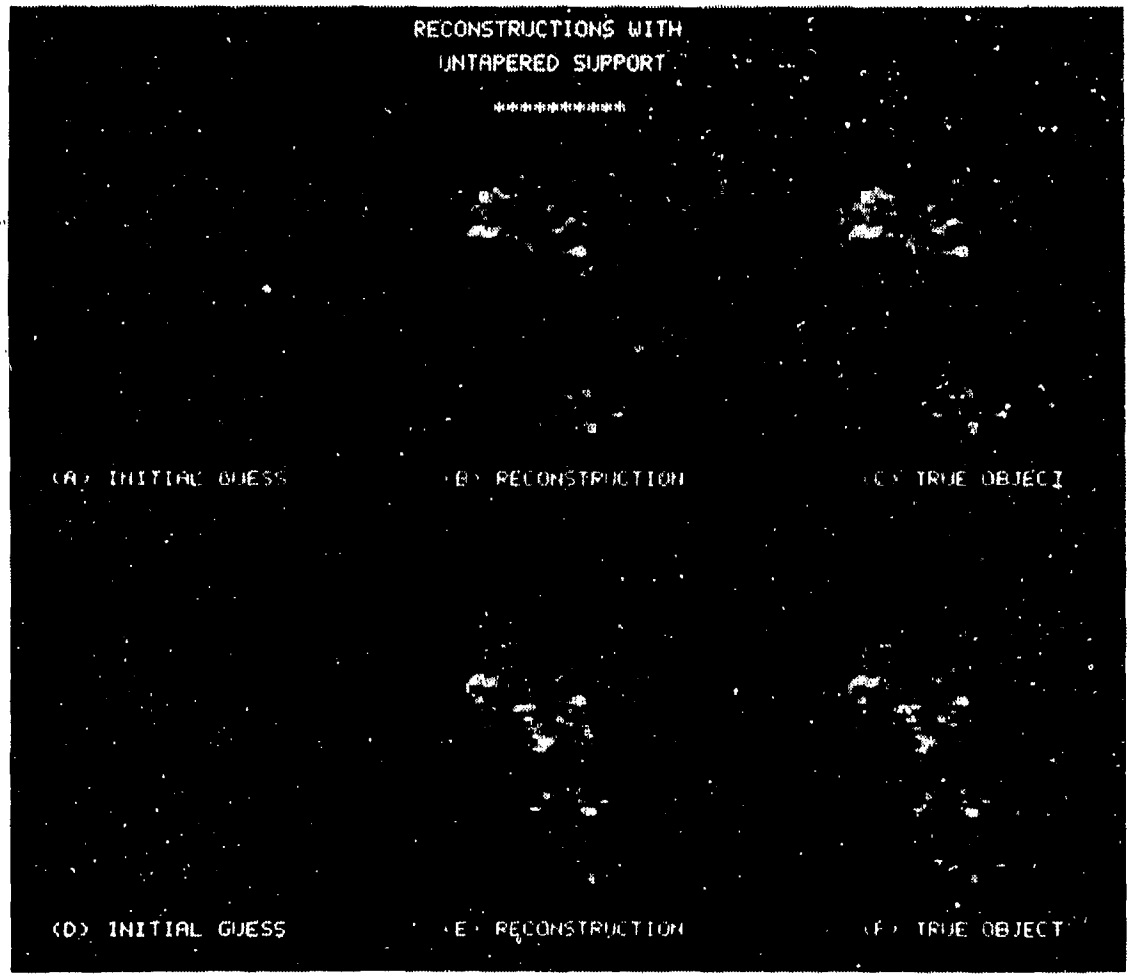


FIGURE 5-5. RECONSTRUCTIONS OF OBJECTS WITH UNTAPERED ILLUMINATION

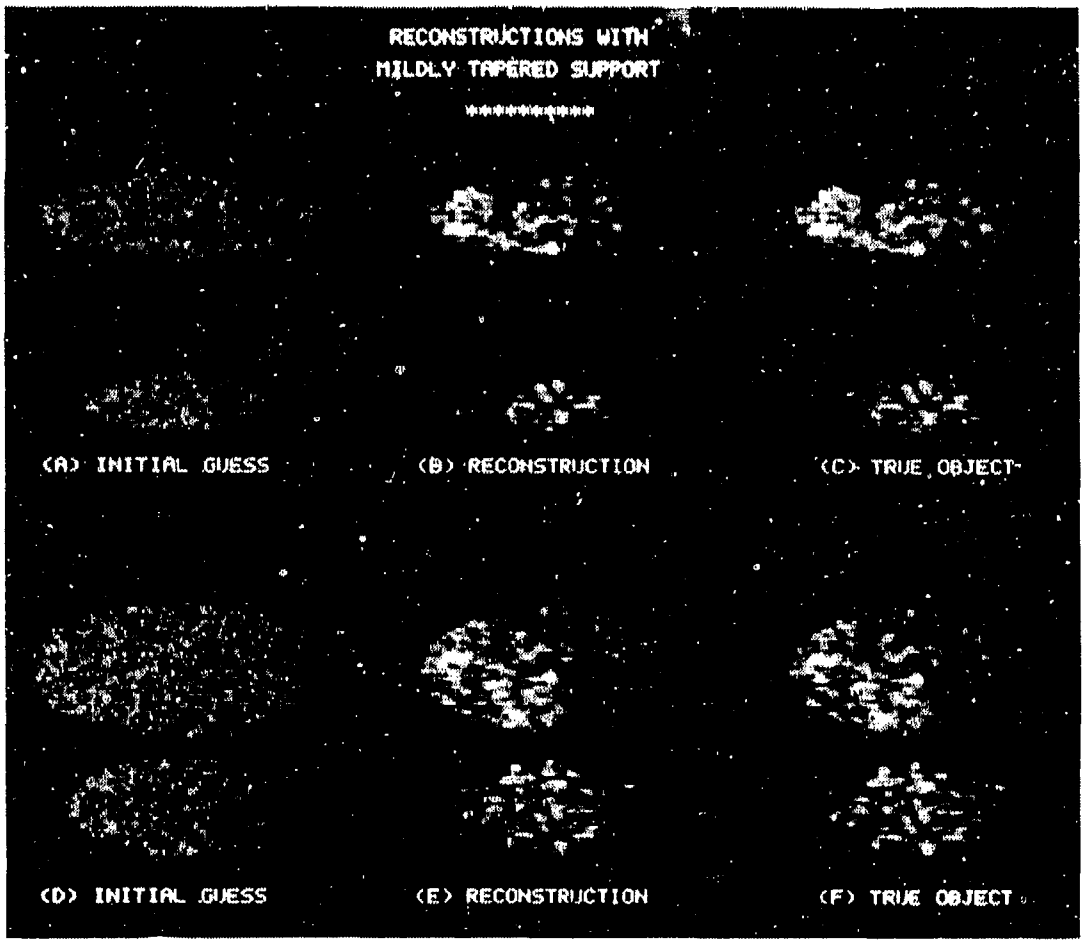


FIGURE 5-6. RECONSTRUCTIONS OF OBJECTS WITH MILDLY TAPERED ILLUMINATION. (Taper #1 in Figure 5-3)

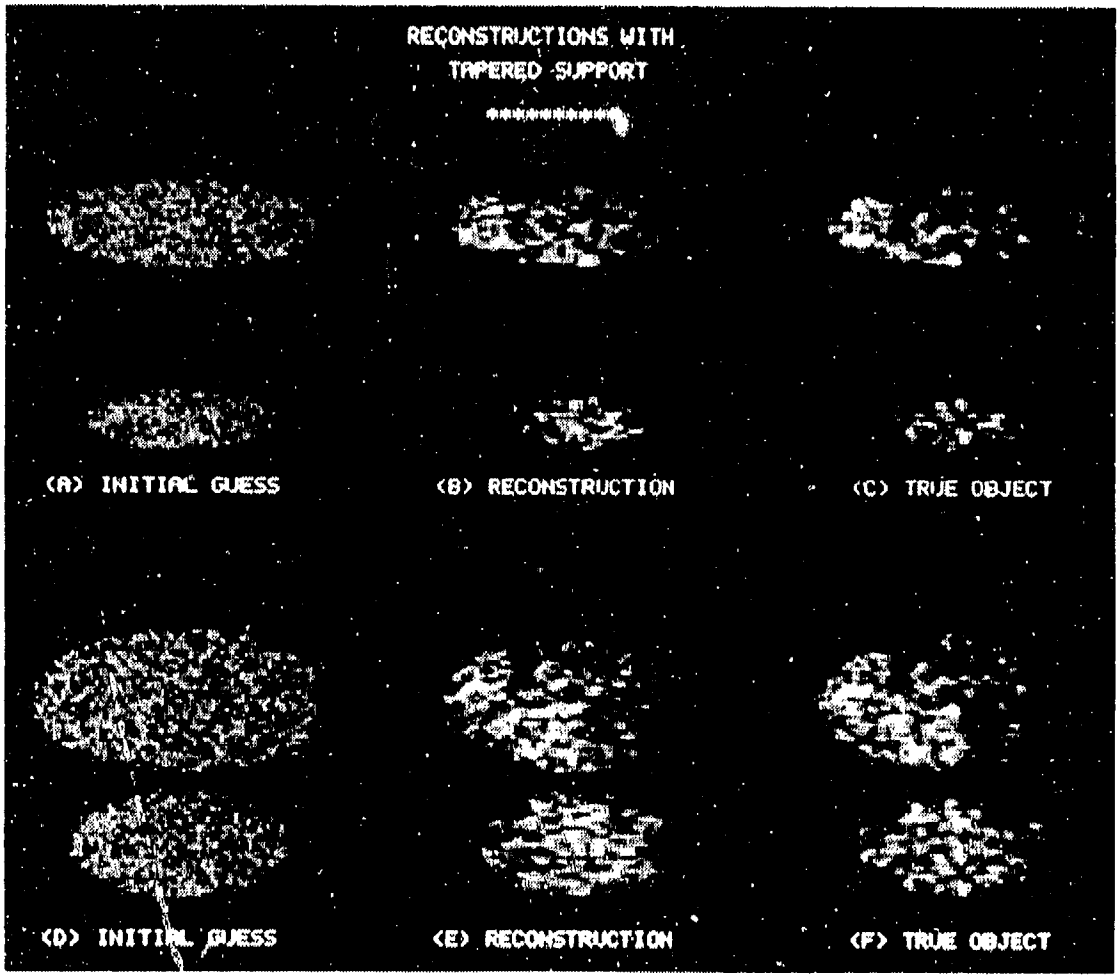


FIGURE 5-7. RECONSTRUCTIONS OF OBJECTS WITH TAPERED ILLUMINATION. (Taper #2 in Figure 5-3)

5-8. Notice that the difference image indicates that the reconstruction is shifted in the horizontal direction relative to the true object. This suggests that the algorithm may be stagnating because of its inability to properly register the reconstruction relative to the support constraint when tapered illumination is used.

To better understand this conjectured mode of stagnation consider an object f(x,y) with tapered illumination. We define a binary mask m(x,y) that is the characteristic function of S, the known support:

$$m(x,y) = \begin{cases} 1, & (x,y) \in S \\ 0, & (x,y) \in S' \end{cases} \tag{5-6}$$

where S' stands for the complement of a S. An image g'(x,y) outputted by the iterative Fourier transform algorithm is the inverse Fourier transform of a Fourier-domain estimate having modulus equal to the given Fourier modulus data coupled with the current estimate of the Fourier phase. Suppose the output image is just a shifted version of the object:

$$g'(x,y) = f(x - x_o, y - y_o). \tag{5-7}$$

A shift in the object domain introduces a linear phase factor in the Fourier domain and has no effect on the Fourier modulus. This output image will clearly satisfy the Fourier modulus constraint. The output image has, however, been shifted relative to the mask so that the object domain support constraint has been violated. In other words, multiplying by the mask function will crop an edge of the output image. We use a normalized error metric to indicate the degree of inconsistency between an estimate and the object support constraint:

$$E_o^2 = \frac{\sum_{x,y} |g'(x,y)m'(x,y)|^2}{\sum_{x,y} |g'(x,y)|^2} \tag{5-8}$$

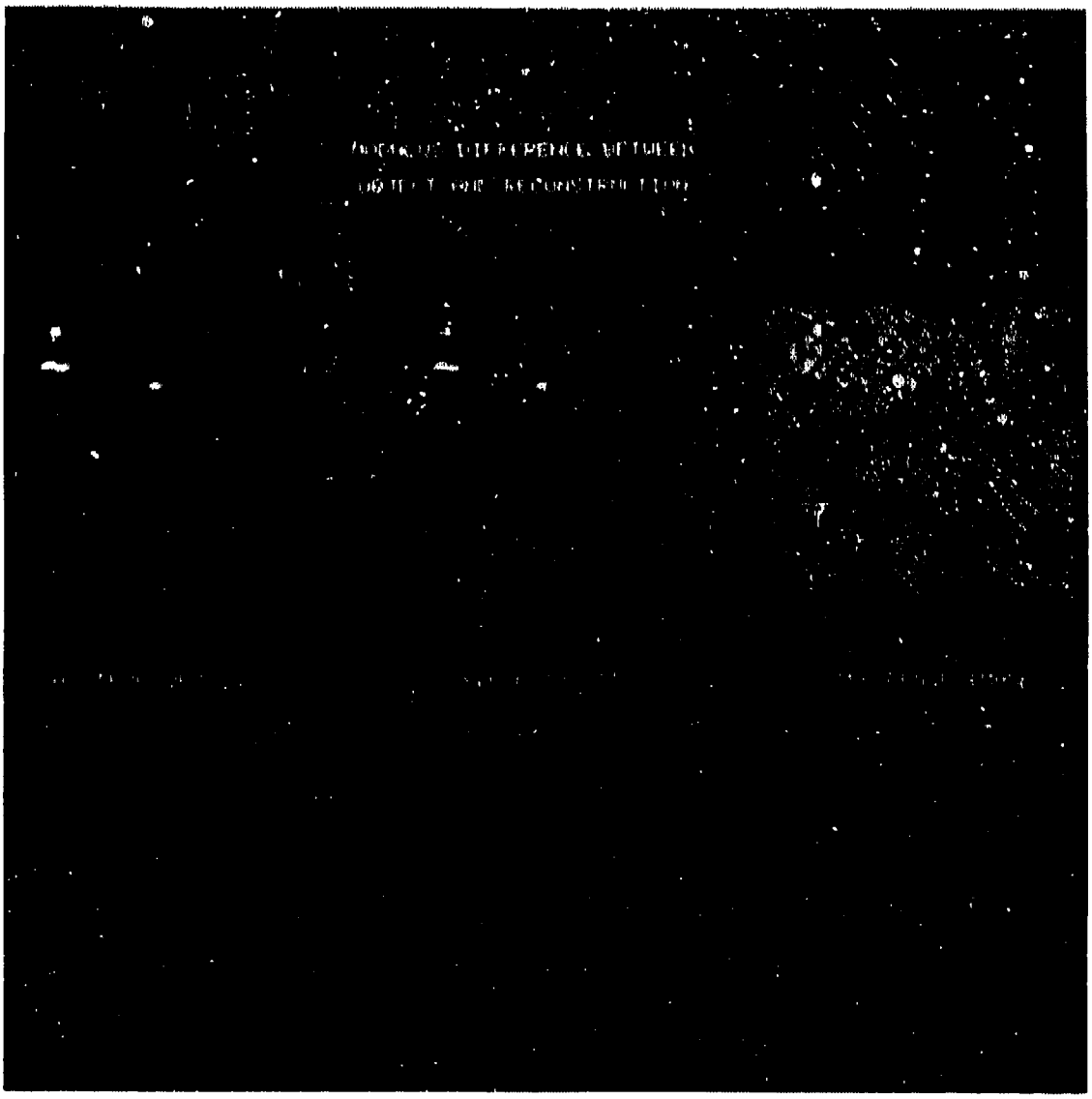


FIGURE 5-8. MODULUS DIFFERENCE BETWEEN OBJECT AND RECONSTRUCTION. (Illumination due to Taper #2 in Figure 5-3) The bipolar difference image has been biased up for display.

where $m'(x,y)$ is the characteristic function of S'. If the shift vector (x_o,y_o) is small with respect to the illumination taper the object domain error metric will also be relatively small. This is because only the tapered edges, where there is little energy, will be cropped and this contributes to only a small portion of the total object energy.

Though the cropped output image now satisfies the support constraint, its Fourier-transform modulus no longer exactly equals $|F(u,v)|$. It can easily be shown that the Fourier-domain error metric is also small. Thus the error metric penalty is small in either domain when shifting a tapered object by a small amount. An algorithm that chooses successive estimates based upon these error metric objective functions will be insensitive to small shifts and would easily stagnate due to extremely small slopes in the objective function. Such an algorithm would be ineffective at finding the proper object registration. Furthermore, one can imagine that, with the right redistribution of the cropped object energy, an object estimate could correspond to a local minimum in the objective function.

Although the mode of stagnation just presented is conjectured, it provides the motivation for the "shrunken-mask" algorithm. The shrunken-mask algorithm is designed to find the proper registration early on in the iterative reconstruction thus circumventing shift-related stagnation that might otherwise appear.

Consider a new binary mask $m_t(x,y)$ created by hardlimiting the tapered illumination function with some intermediate threshold value:

$$m_t(x,y) = \begin{cases} 1, & (x,y) \text{ such that } |w(x,y)| > t \\ 0, & (x,y) \text{ such that } |w(x,y)| \leq t \end{cases} \tag{5-9}$$

where t is the threshold value, $0 \leq t \leq 1$. Notice that $m_t(x,y)$ will be a "shrunken" version of the full mask m(x,y) defined for t = 0. Suppose that we employ the shrunken mask as the support constraint. If we crop the true object with the shrunken mask this will yield an estimate with a modest penalty in both the object and Fourier domains, so long as the threshold value is not too large. Notice, however, that a shift in this cropped estimate will yield an object-domain penalty, due to the shrunken mask and the artificially created discontinuous object edges, that is much greater than the penalty that would be due to the normal support constraint. Thus we would expect the output image to be centered better with the shrunken mask.

While the Fourier modulus and the shrunken-mask support constraints are inconsistent, they may still be jointly enforced in an iterative reconstruction algorithm to get an intermediate reconstruction. We might expect this intermediate result to display gross features of the true object in proper registration. Enlarging the mask to its full size (setting t = 0) removes the constraint inconsistency and allows for a complete reconstruction that hopefully avoids shift-related local minima. The shrunken-mask algorithm is shown schematically in Figure 5-9.

The shrunken-mask algorithm was first tested on the elliptical object where the taper (taper #2 in Figure 5-3) induced stagnation in previous trials. The convergence characteristics are displayed in Figure 5-10. It is clear that the conventional algorithm performed better early on in the iterative sequence. This is reasonable since the support constraint is initially looser and easier to satisfy. By contrast, the shrunken-mask algorithm error metric quickly levels off while an intermediate reconstruction is being produced but drops dramatically when the full-size mask is introduced.

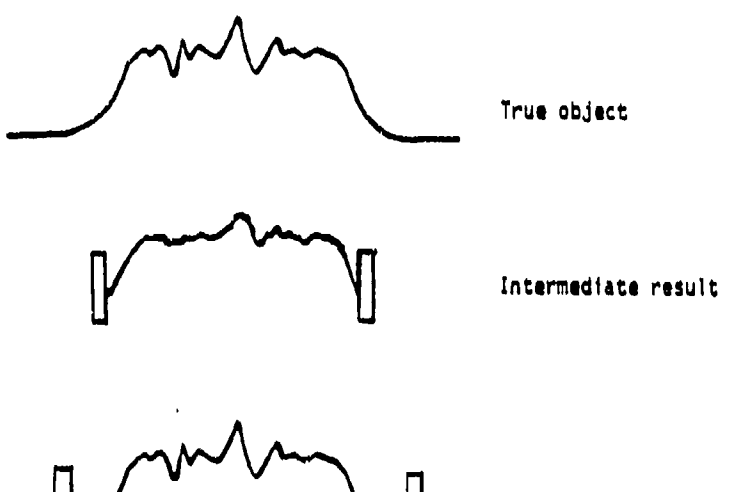


FIGURE 5-9. THE SHRUNKEN-MASK ALGORITHM

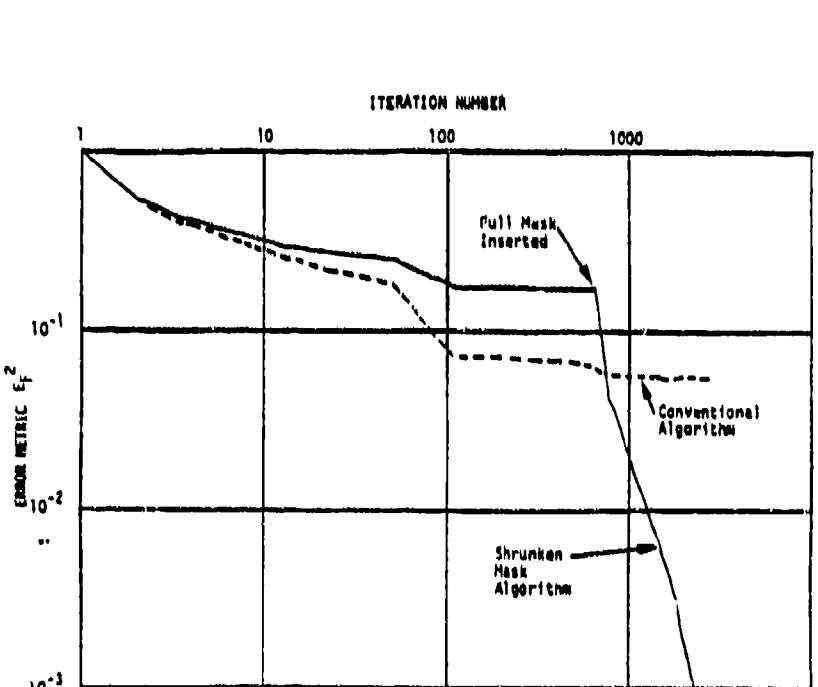


FIGURE 5-10. CONVERGENCE FOR SHRUNKEN-MASK ALGORITHM. Taper is Taper #2 in Figure 5-3. The shrunken mask had a threshold value t = .9.

5.4 THE ENLARGING MASK ALGORITHM

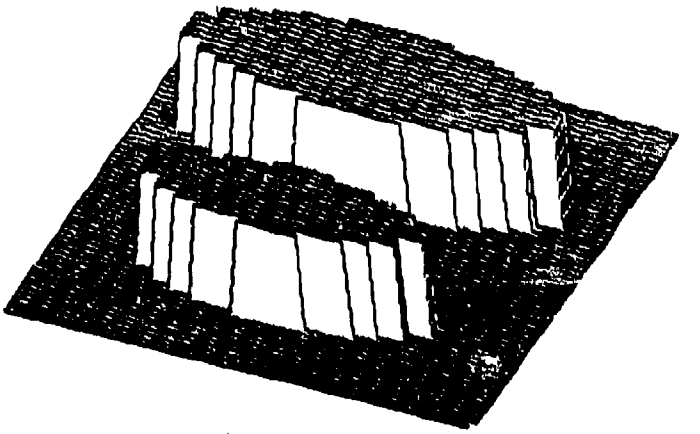
While the success in the shrunken-mask algorithm is encouraging, the amount of illumination taper for which it worked remains extremely small. A much more substantial taper was introduced by using a circular convolution kernel with a radius of 4 pixels. The resultant illumination pattern is shown in Figure 5-11. When the shrunken-mask algorithm was applied to an object with this illumination, the convergence was not much better than the conventional algorithm. This was true for a variety of threshold values that were tested. Apparently the increased taper is a significant obstacle for the shrunken-mask algorithm.

Recall that the shrunken-mask algorithm jumps from a small mask to the full mask in a single step. A logical generalization of the shrunken-mask algorithm uses several intermediate-size masks in order to make a more gradual transition to the full size mask. We call this the "enlarging-mask" algorithm. The collection of masks used in a given application is characterized by a sequence of threshold values. The convergence curve for the enlarging-mask algorithm, when applied to an object with this increased taper, is shown in Figure 5-12(a). The scallop effect exhibited by the convergence curve is due to the successive application of increasingly enlarged masks. The enlarging-mask algorithm clearly out-performs the shrunken-mask algorithm and the final reconstruction exhibits very good agreement with the data and support constraint.

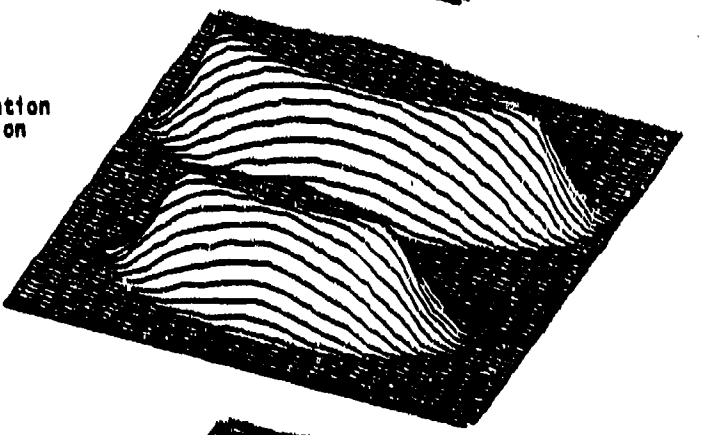
A final trial was performed with an even more realistic illumination taper created with a Gaussian-like convolution kernel with a maximum radius of 6 pixels. This kernel, K(r), was formed by correlating a circle function with a radius of 2 pixels with its own autocorrelation:

$$K(r) = \mathrm{CIRC}\,(r/2) ** \mathrm{CIRC}\,(r/2) ** \mathrm{CIRC}\,(r/2), \qquad (5\text{-}10)$$

A. No taper



B. Taper due to convolution with a circle function (radius = 4 pixels)



C. Taper due to convolution with a Gaussian-like function (max radius = 6 pixels)

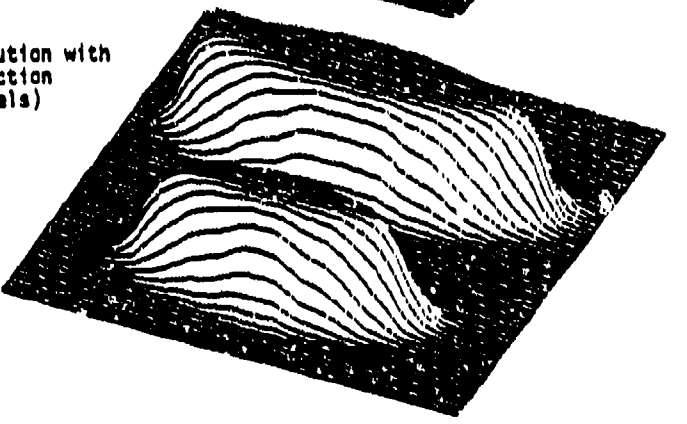


FIGURE 5-11. ILLUMINATION PATTERNS

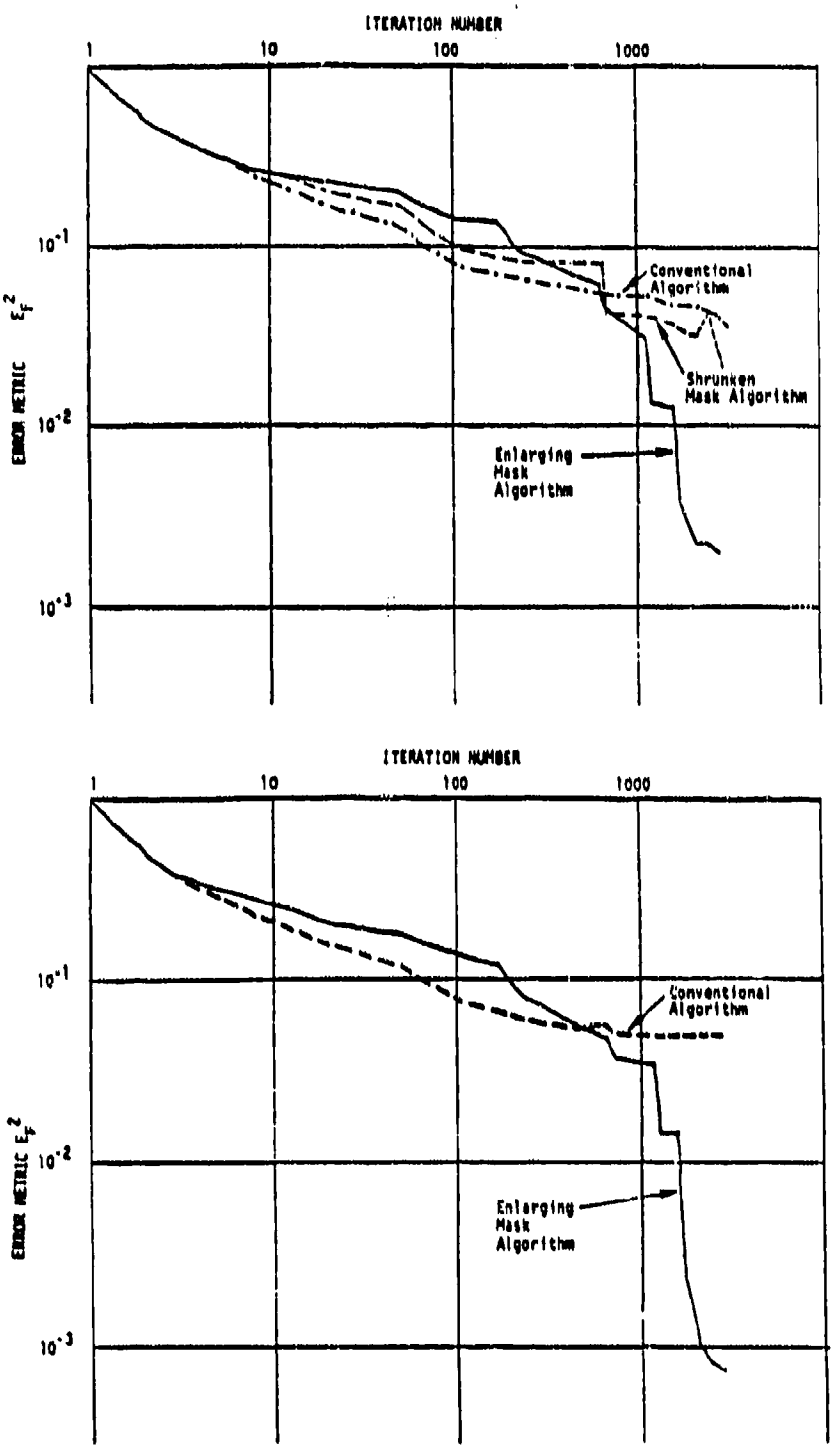


FIGURE 5-12. COMPARISON OF CONVERGENCE BEHAVIOR OF THREE ALGORITHMS. For illumination taper due to (a) a circular convolution kernel of radius 4 pixels, and (b) a Gaussian-like convolution kernel of maximum radius 6 pixels. The threshold sequence for the enlarging-mask algorithm was 0.5, 0.3, 0.2, 0.1, 0.001.

where the double star indicates two-dimensional cross-correlation. This kernel is a close approximation to a two-dimensional Gaussian function. The resultant illumination pattern is shown in Figure 5-11. Note that this illumination has a smoother taper and that the tails extend out further at very low levels. The convergence curves for this case are shown in Figure 5-12(b). Again the enlarging-mask algorithm succeeds at finding a reconstruction that is in excellent agreement with the data and support constraint whereas the conventional algorithm did not. This reconstruction is visibly indistinguishable from the true object. The results of reconstructions performed with and without the enlarging-mask algorithm are given in Figure 5-13 for illumination patterns due to the circular and Gaussian convolution kernels.

While tapered illumination presents significant stagnation problems for conventional phase-retrieval algorithms, these examples demonstrate that the enlarging-mask algorithm successfully circumvents these difficulties, even in the presence of large amounts of taper.

The success of the enlarging-mask algorithm leads us to ask how much taper can be introduced before the performance of the algorithm deteriorates. In order to investigate this question, a new series of simulations was performed using a frame size of 256 x 256 pixels. The source of the illumination pattern was changed from the pair of ellipses to an isosceles triangle with 50 pixels on a side. These choices were made for multiple reasons. In the first place, this simulation design allows for the introduction of extreme amounts of taper without aliasing problems. Secondly, the disjoint elliptical patterns were dropped since a single connected illumination pattern results when enough taper is added. Because of this, it would be difficult to determine if a decrease in performance were due to increased taper, the breakdown in the disjoint nature of the illumination pattern, or a combination of the two. The use of a single connected illumination pattern helps to isolate the effects due to taper alone. Finally, the triangular shape

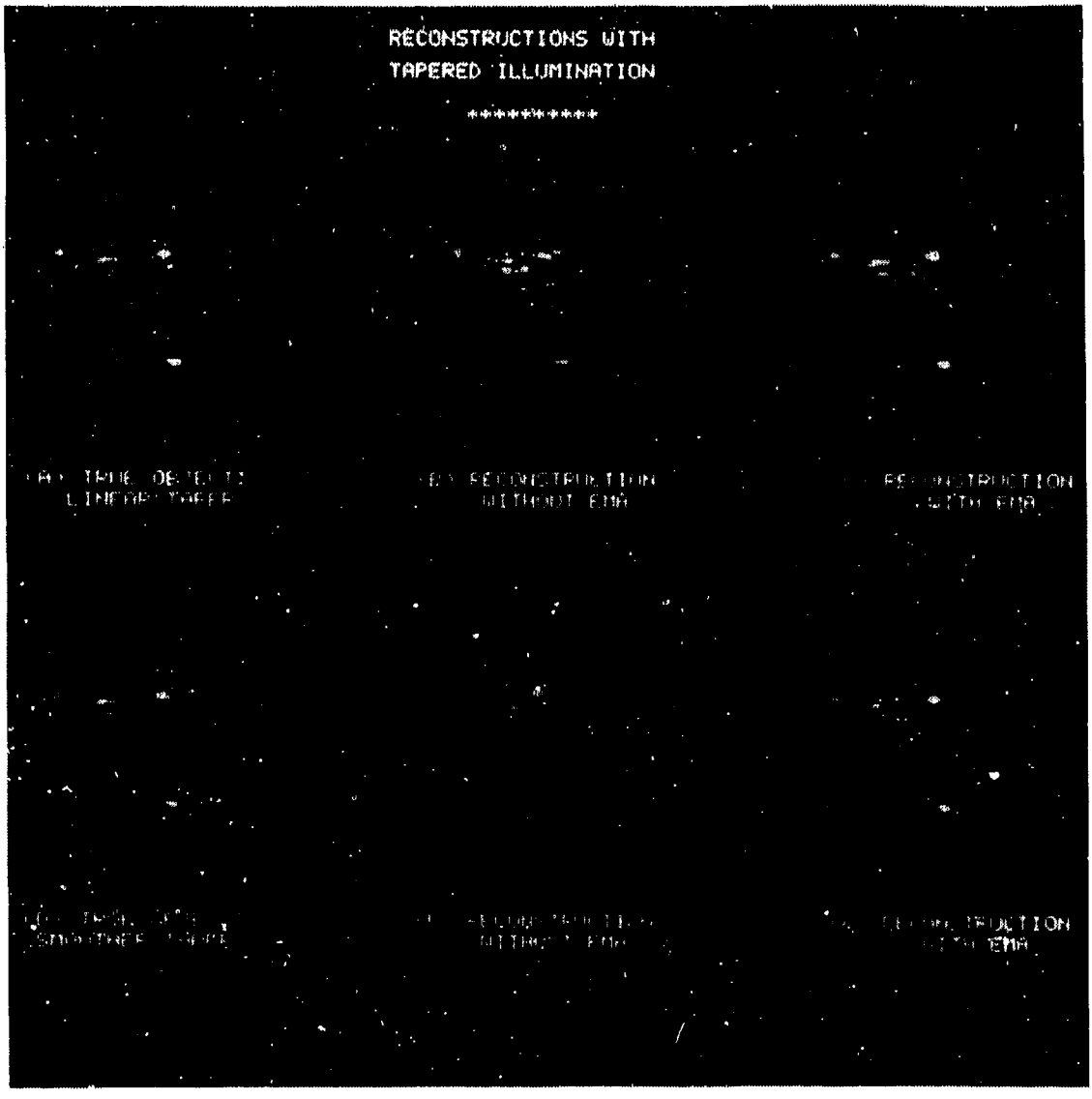


FIGURE 5-13. RECONSTRUCTIONS WITH AND WITHOUT THE ENLARGING-MASK ALGORITHM (EMA). The illumination pattern in A-C is shown in Figure 5-11B. The illumination pattern in D-F is shown in Figure 5-11C.

was selected because it provides the best known support constraint for phase retrieval (see Section 4.1).

In a real system, taper will unavoidably be introduced by convolving a (possibly binary) source illumination pattern, b(x), with a system impulse response (IPR) function, a(x):

$$w(x) = b(x) * a(x) \quad , \tag{5-11}$$

where w(x) is the coherent illumination pattern, as before. In an optical system the IPR is the Fourier transform of the coherent transfer function or aperture function, A(u). The casting of an illumination pattern by a coherent optical system is illustrated in Figure 5-14. Similar concepts apply to radar systems.

In order to minimize the amount of energy that is diffracted to regions far afield of the desired binary illumination pattern, the side lobes of the IPR should be reduced. This is done conventionally with an apodising aperture. The form of the particular apodization used in these simulations is given here:

$$\begin{aligned} A(u) &= A(\rho) \\ &= \begin{cases} \left[1 - \left(\frac{\rho}{\rho_c}\right)^2 \right]^2 & \rho \leq \rho_c \\ 0 & \rho > \rho_c \end{cases} \end{aligned} \tag{5-12}$$

where $\rho = |u|$ and ρ_c represents the cutoff frequency. Figure 5-15 shows a cross section of such an apodising aperture and the apodised IPR that results from it. This particular apodising function is one of many used conventionally [5.4] although it is not necessarily optimal for the enlarging-mask algorithm.

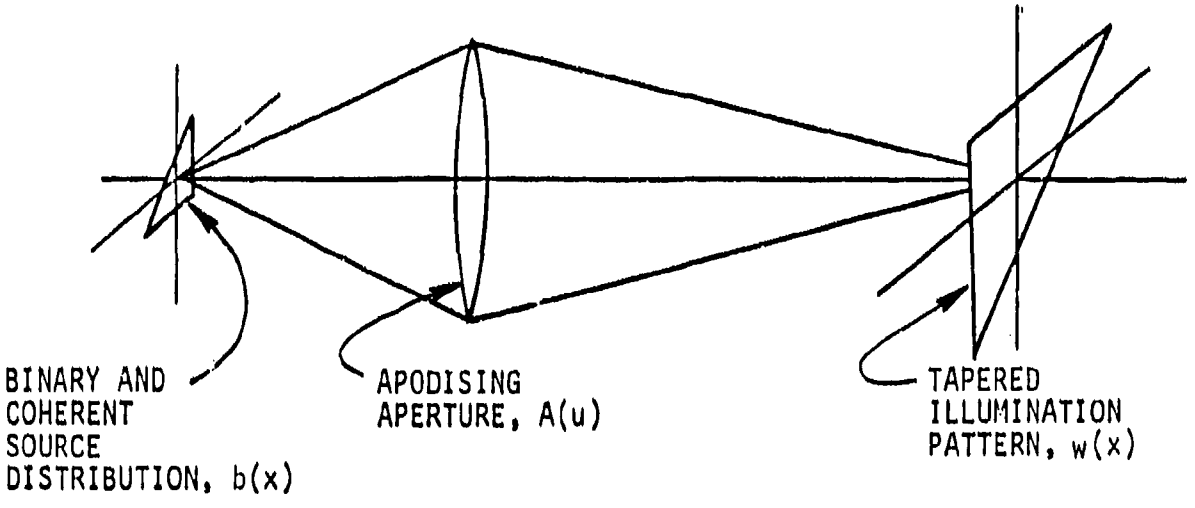


FIGURE 5-14. FORMATION OF TAPERED ILLUMINATION PATTERN

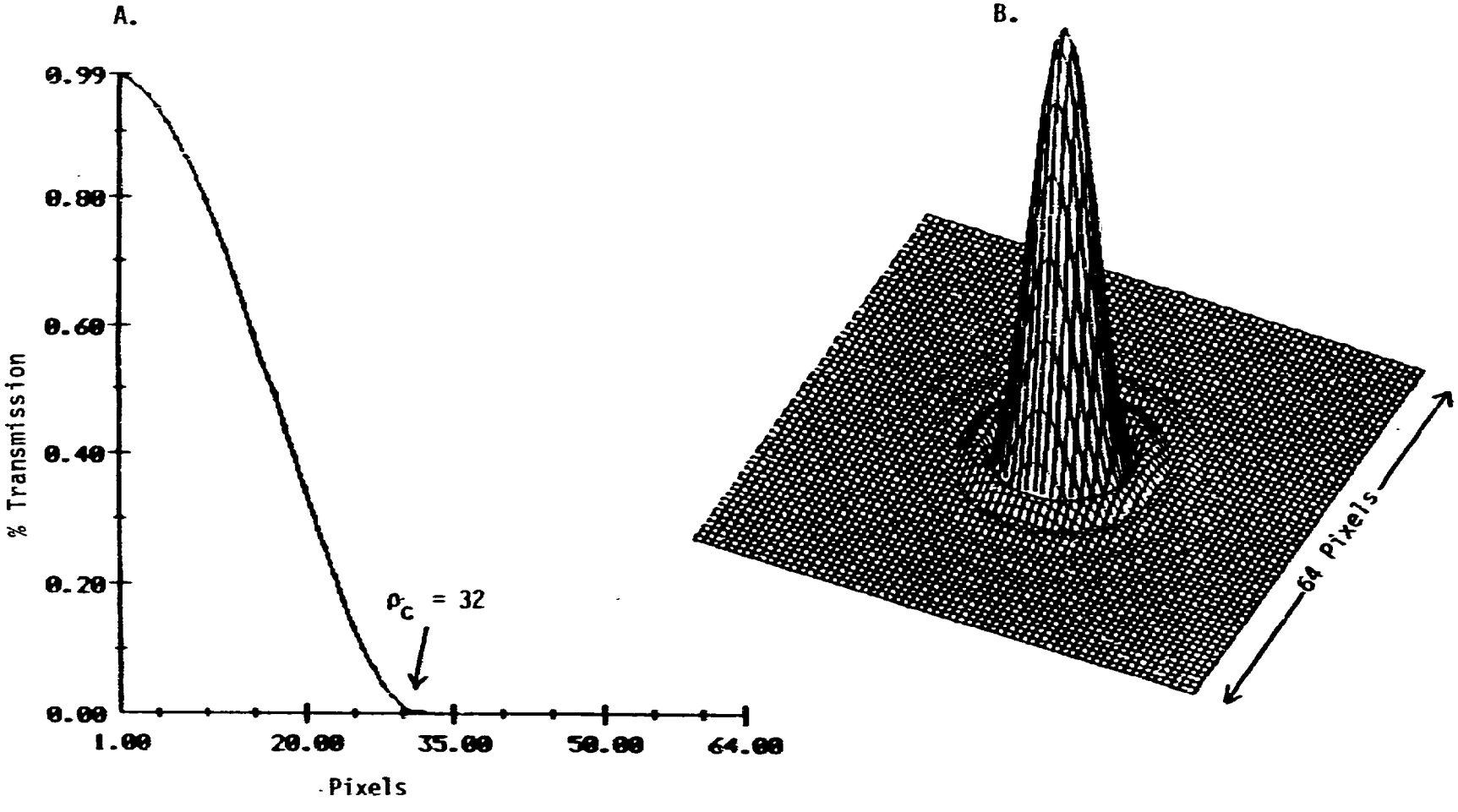


FIGURE 5-15. APODISING APERTURE AND IMPULSE RESPONSE FUNCTION.
A. Cross section of circularly symmetric apodising function.
B. Intensity plot of resulting impulse response function.

Four illumination patterns were produced, each with a different amount of taper. The amount of taper introduced was controlled by varying the value of the cutoff frequency in Eq. (5-12). The four illumination patterns are shown in Figure 5-16. These illumination patterns were applied to actual SAR data of an extended cultural scene to get four effective objects. The Fourier intensity for each effective object represents the raw data in this simulation. The Fourier modulus for each effective object is shown in Figure 5-17. A close examination of Figure 5-17 reveals that the Fourier modulus does, in fact, change with varying amounts of taper. It is difficult to see any specific trends in the Fourier modulus that is indicative of the amount of taper present.

A standardized enlarging-mask algorithm was exercised with each data set. The details of this algorithm are shown in Table 5-1. While the choice of number of iterations at each threshold is probably reasonably good, we do not claim that it is optimum.

Table 5-1
STANDARDIZED ENLARGING-MASK ALGORITHM

Threshold (% of Illumination Peak)	Number of Iterations
90	150
70	200
50	200
30	200
10	700
1	400

The effective or true objects with varying amounts of taper and their associated enlarging-mask reconstructions are displayed in Figure 5-18. These reconstructions indicate that for very large amounts of taper (12-16 pixels) even the enlarging-mask algorithm is unable to produce a

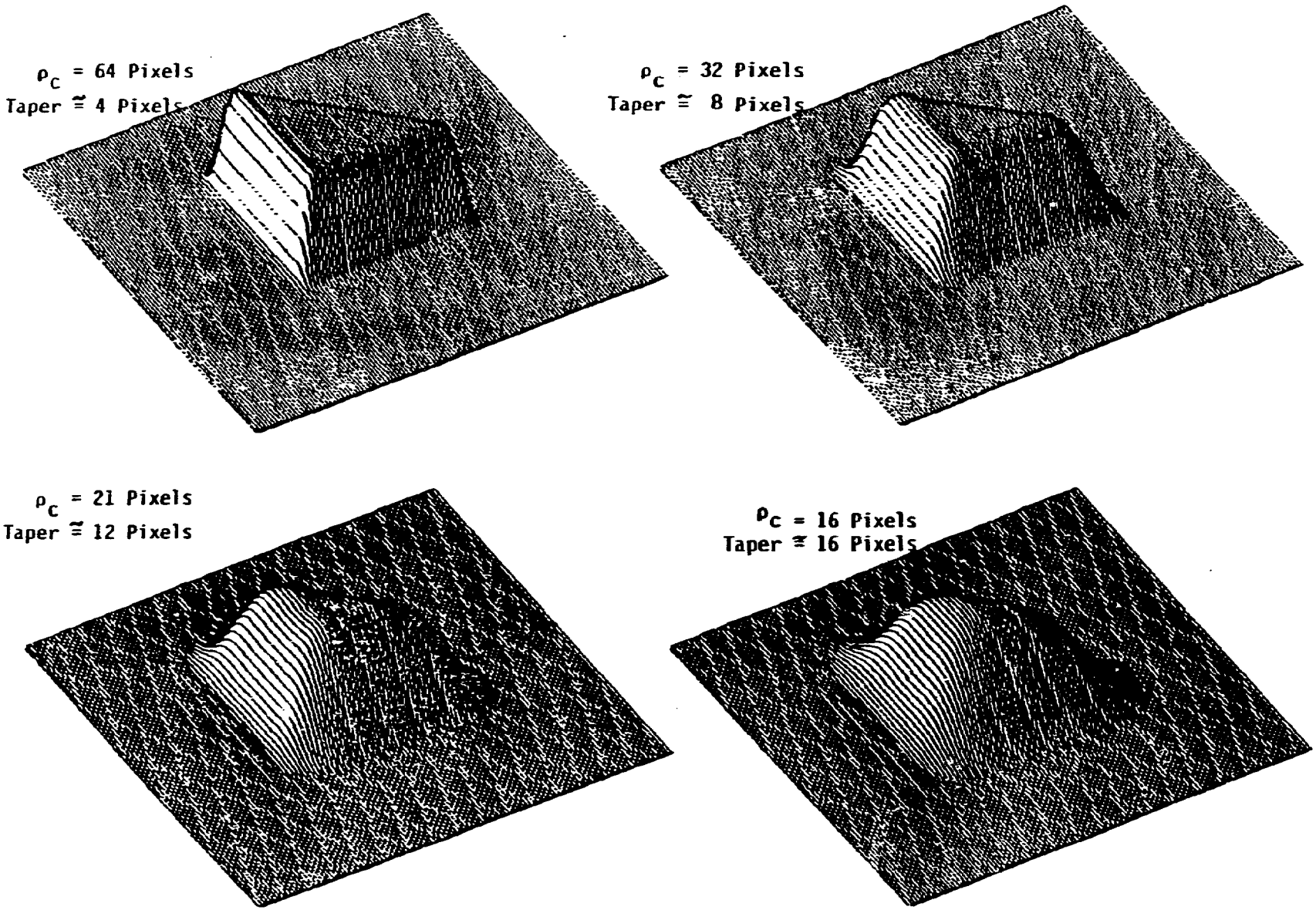


FIGURE 5-16. ILLUMINATION PATTERNS WITH VARYING AMOUNTS OF TAPER.

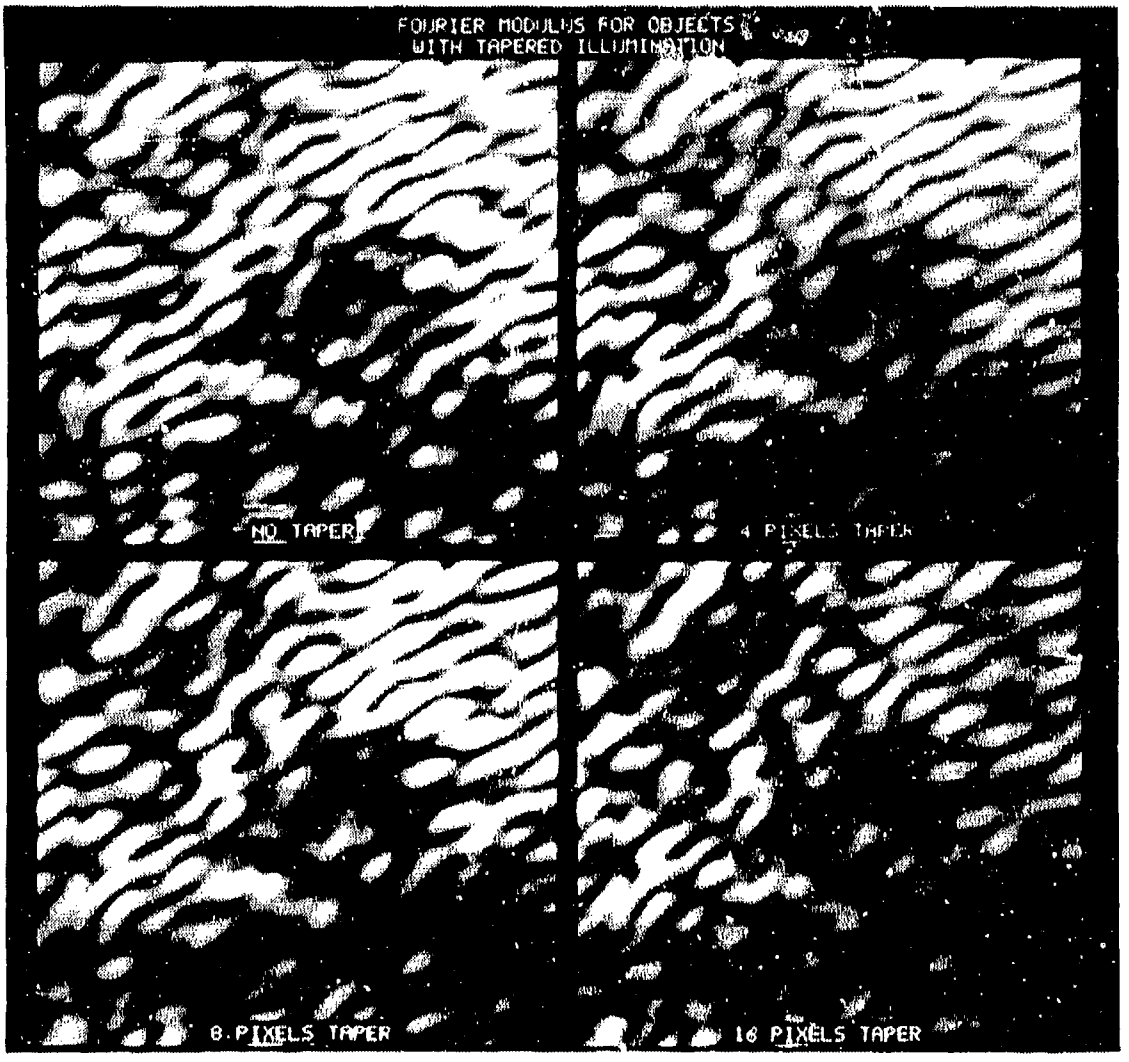


FIGURE 5-17. FOURIER MODULUS FOR OBJECTS WITH TAPERED ILLUMINATION.

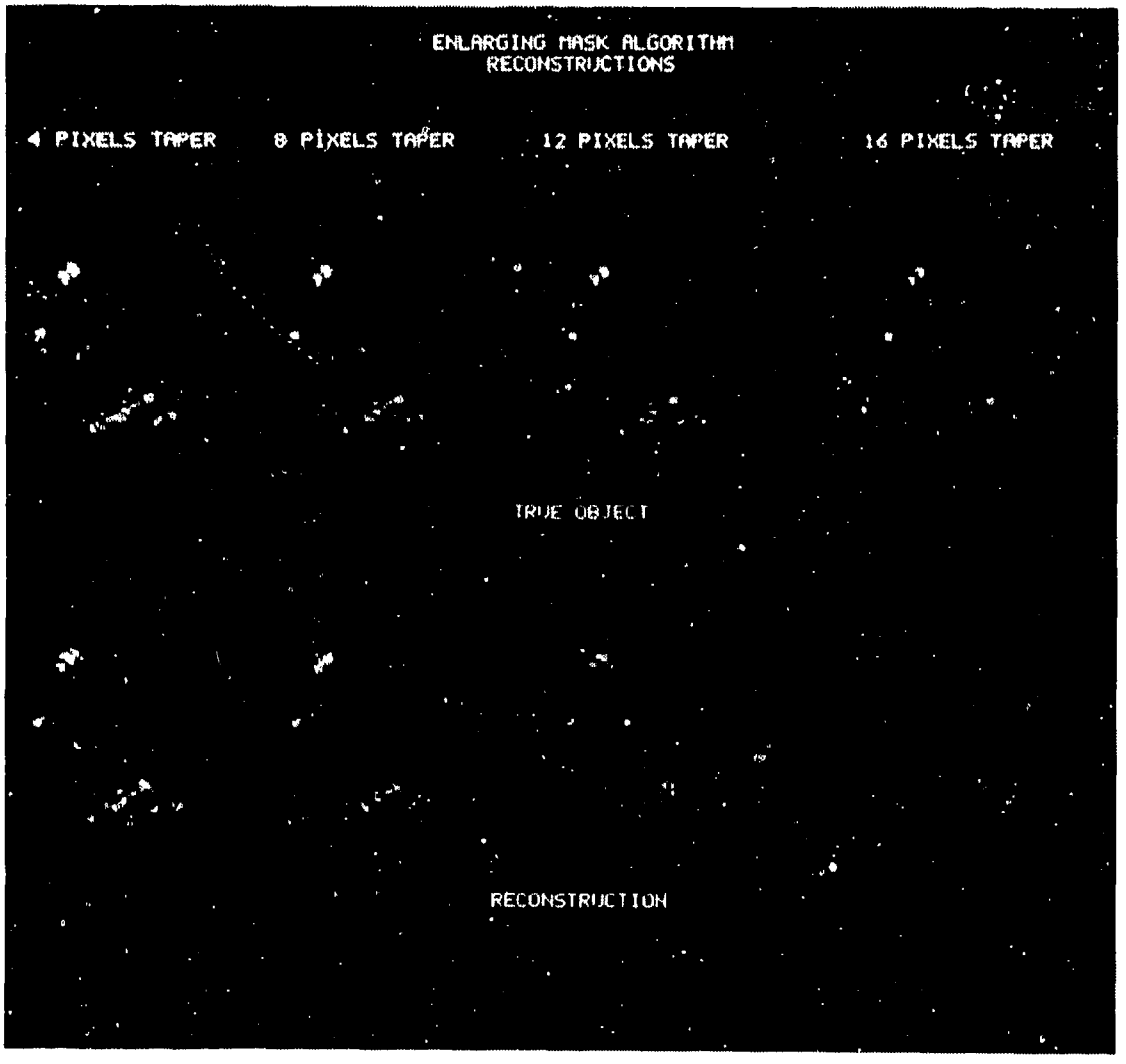


FIGURE 5-18. TRUE OBJECTS AND ENLARGING-MASK RECONSTRUCTIONS FOR VARYING AMOUNTS OF TAPER.

high-fidelity reconstruction. Nevertheless, low-frequency information is restored. An overexposed version of the images in Figure 5-18 is given in Figure 5-19. This illustrates that the true object with 16 pixels of taper really does have energy at the extremes of the tapered illumination. In addition, it is clear that the enlarging-mask algorithm tries to reconstruct too much energy in the taper region.

It is useful to define a mean-square error figure of merit between the reconstruction and the true object. Of course such a figure of merit would not be available in a true imaging application because the true object would not be available for the computation. Nevertheless, this figure of merit is very useful in characterizing the performance of the algorithm in controlled settings such as computer simulations. The normalized absolute error is defined:

$$E_g = \left[\frac{\sum_x |ag(x-x_o) - f(x)|^2}{\sum_x |f(x)|^2} \right]^{1/2}, \qquad (5\text{-}13)$$

where $f(x)$ is the true object and $g(x)$ is the reconstruction. Because phase retrieval may be unable to provide absolute registration of the reconstruction, $g(x)$ is optimally shifted by the vector x_o prior to the computation of E_g. Furthermore, the reconstruction will always have a constant phase ambiguity. Therefore the complex coefficient a is also selected to minimize E_g. The normalized absolute error is plotted as a function of number of pixels of taper in Figure 5-20.

5.5 The Effects of Noise and Taper

In order to provide guidance for the design of a real system one needs to know what the combined effect of tapered illumination and noisy data is on the quality of reconstruction. The first experiment

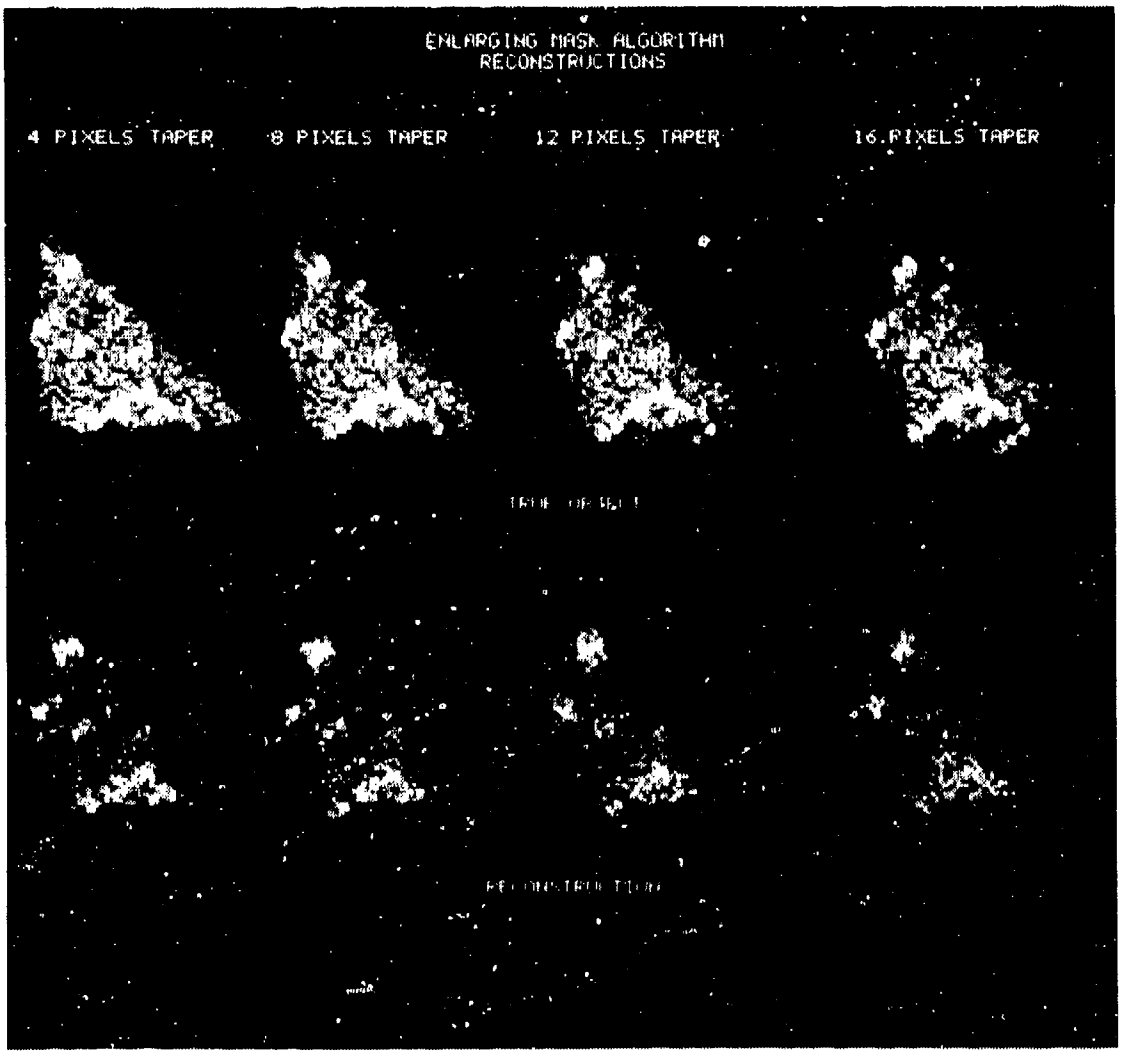


FIGURE 5-19. OVEREXPOSED OBJECTS AND ENLARGING MASK RECONSTRUCTIONS.

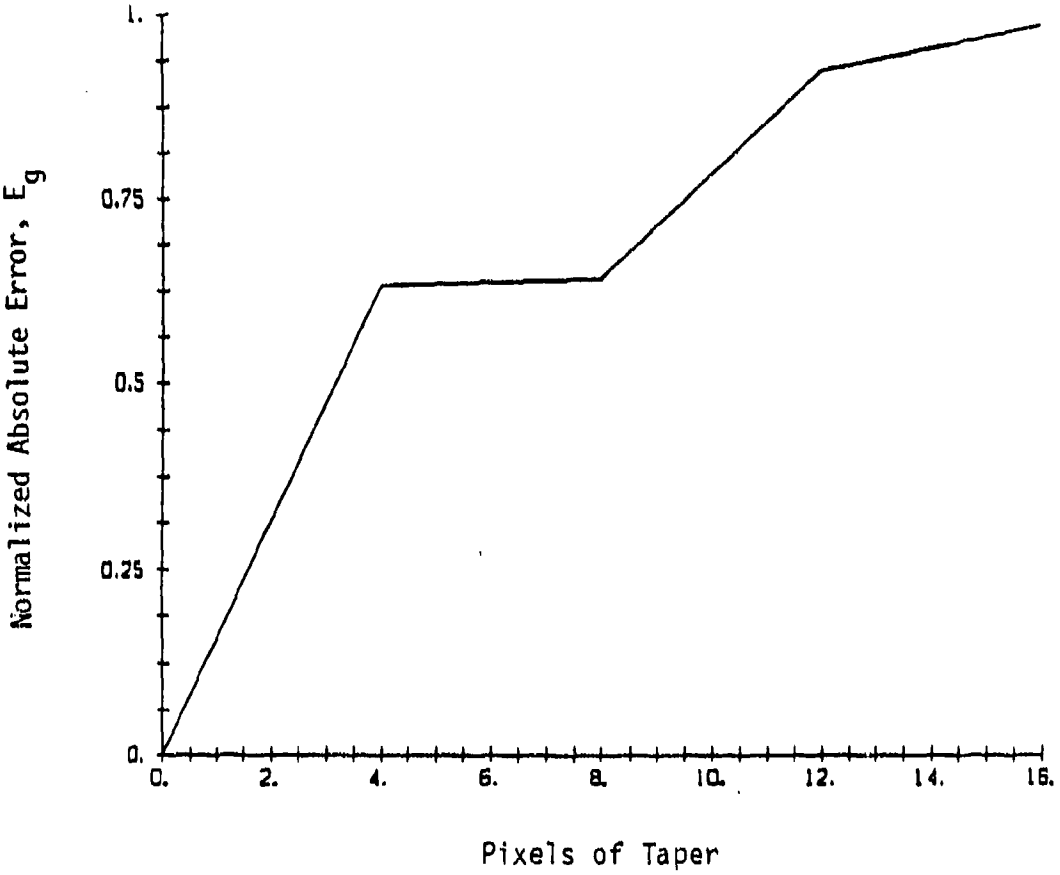


FIGURE 5-20. NORMALIZED ABSOLUTE ERROR IN RECONSTRUCTION AS A FUNCTION OF AMOUNT OF TAPER.

performed to explore this issue utilized an untapered triangular illumination pattern, and varying amounts of zero-mean Gaussian noise were added to the Fourier intensity. Any negative values that resulted were set to zero, giving an estimated Fourier intensity

$$|\hat{F}(u)|^2 = \begin{cases} |F(u)|^2 + n(u) & , \quad |F(u)|^2 + n(u) > 0 \\ 0 & , \quad \text{otherwise} \end{cases} \qquad (5\text{-}14)$$

where n(u) is one realization of the random noise process with a standard deviation of σ_n. The amount of noise is quantified by percent noise, defined here as the ratio of the standard deviation of the noise to the mean value of the Fourier intensity

$$\% \text{ Noise} = \sigma_n^{-1} \frac{1}{N} \sum_u |F(u)|^2$$

$$= \frac{\sigma_n}{I} \quad , \qquad (5\text{-}15)$$

where N is the number of pixels in the Fourier intensity and I is the mean Fourier intensity. Under the assumption that the noiseless Fourier intensity is a fully-developed speckle pattern, the intensity will have negative-exponential statistics. As a result the standard deviation in the intensity, σ_I, is equal to the mean intensity and Eq. (5-15) can be rewritten

$$\% \text{ Noise} = \frac{\sigma_n}{\sigma_I}$$

or

$$\frac{\sigma_n^2}{\sigma_I^2} = (\% \text{ noise})^2. \qquad (5\text{-}16)$$

The ratio of the variance of the noise to that of the signal, as expressed in Eq. (5-16), is an alternative way of quantifying the amount of noise present.

An estimate of the Fourier modulus is produced by taking the square root of the estimated Fourier intensity. The estimated Fourier modulus is used to enforce the Fourier constraint in the iterative phase-retrieval algorithm. A one-dimensional cross section of the estimated Fourier modulus is given in Figure 5-21 for the cases of no noise and 10% noise. A small area of the estimated Fourier modulus is presented in Figure 5-22 with varying amount of noise. The effect of the noise is more pronounced near the nulls in the Fourier modulus, as expected.

The iterative phase retrieval algorithm was exercised using the triangular support and the estimated Fourier modulus with varying amounts of noise. The number of iterations and the ordering of the hybrid input-output and the error reduction iterations were the same that were used in the earlier enlarging mask experiments (Table 5-1). Figure 5-23 presents reconstructions for the various noise levels and the true object is given for comparison. The quality of the reconstructions is excellent up to the 10% noise level. At noise levels of 15% and higher the reconstructions are noticeably degraded. Nevertheless the gross features of the object persist even at the 30% noise level. These results confirm our view that the phase-retrieval problem is not ill-conditioned, or hypersensitive to noise. Rather, as moderate levels of noise are added in increasing amounts to the raw data, the respective reconstructions degrade gracefully.

The Fourier-domain error metric, Eq. (5-8), associated with the final reconstruction is plotted as a function of % noise in Figure 5-24. Five reconstruction sequences were performed for the case of 5% noise, each with a different noise realization. It is unclear why these

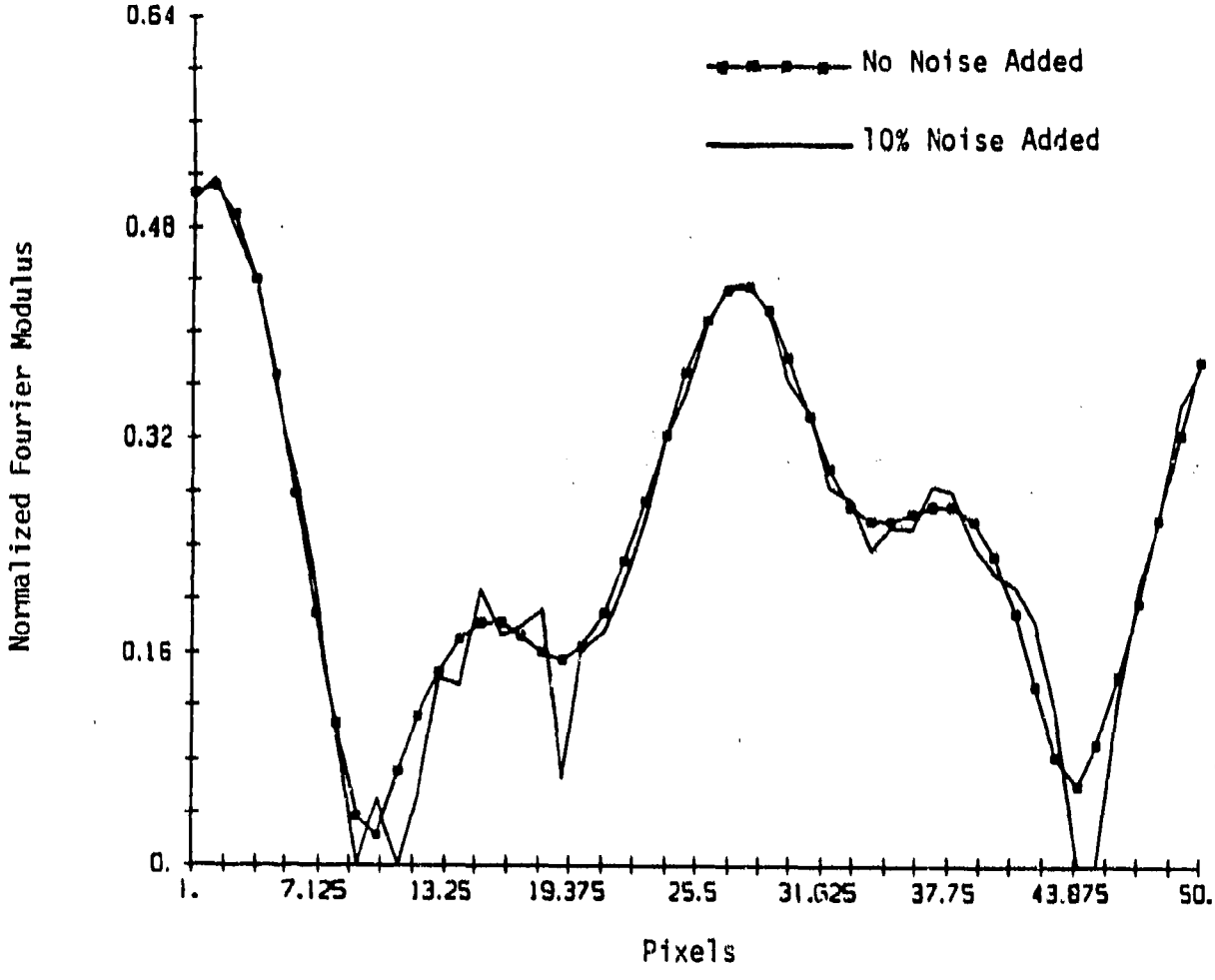


FIGURE 5-21. CROSS SECTION OF NORMALIZED FOURIER MODULUS, WITH AND WITHOUT NOISE.

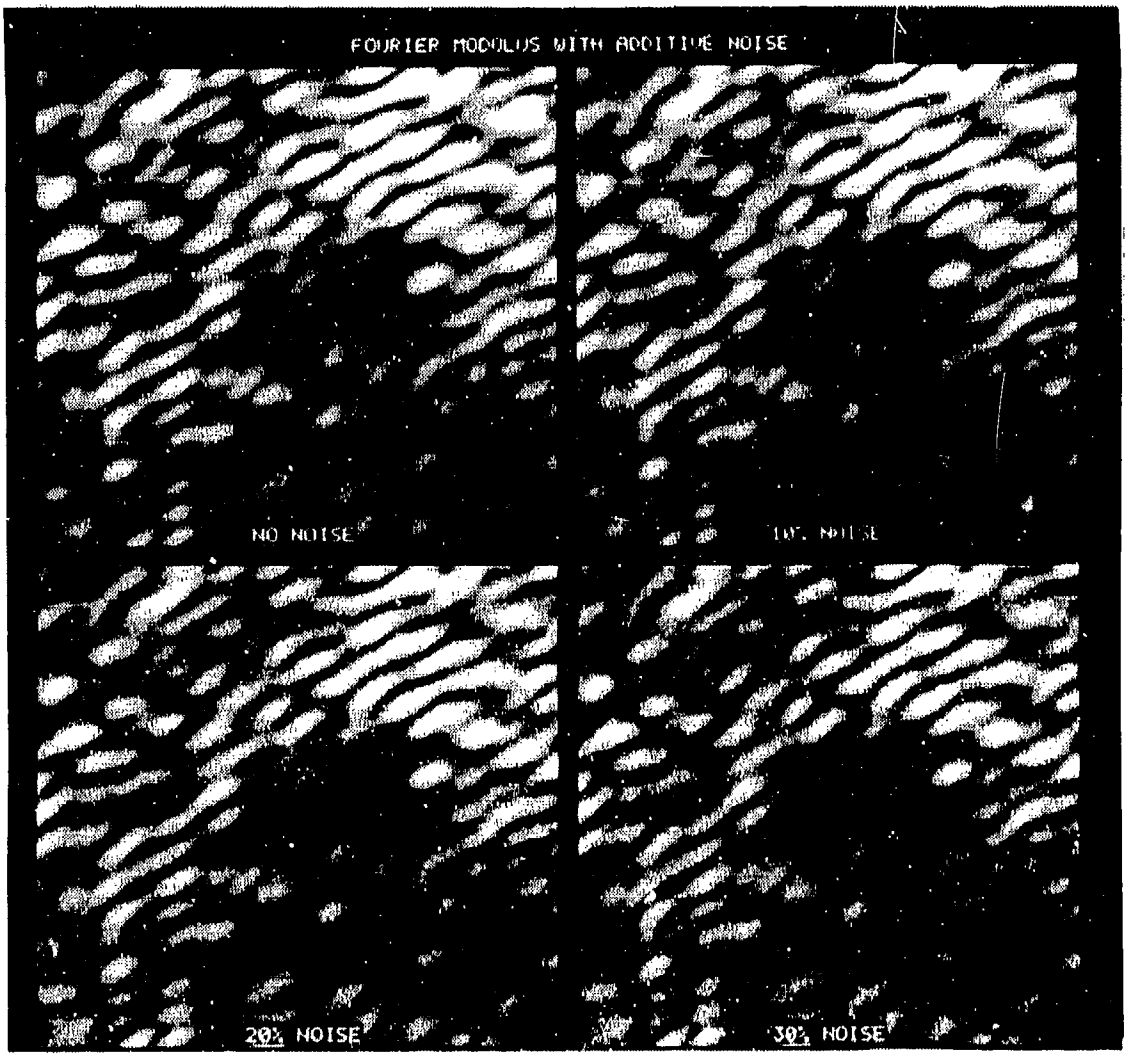


FIGURE 5-22. FOURIER MODULUS WITH VARYING AMOUNTS OF ADDITIVE NOISE.

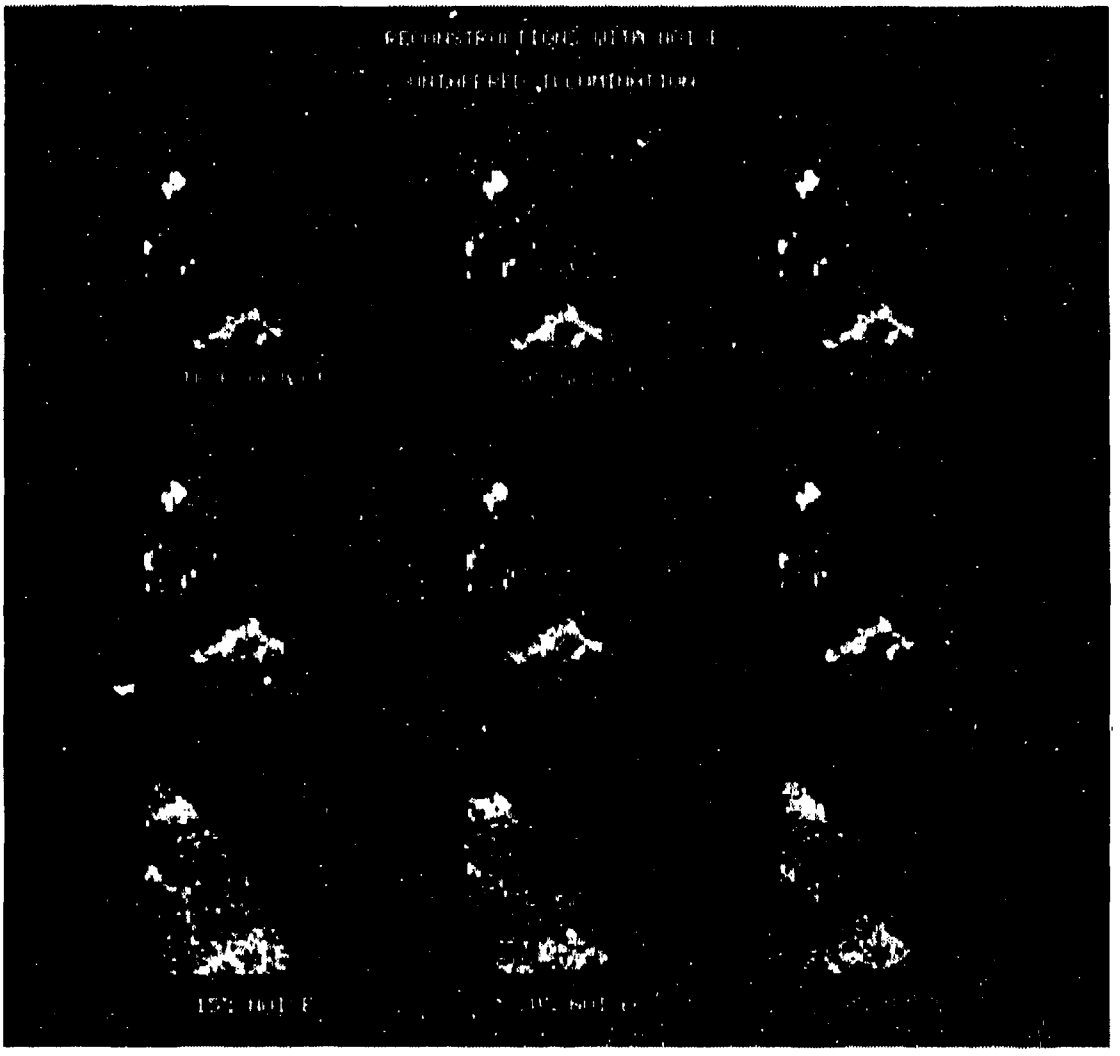


FIGURE 5-23. RECONSTRUCTIONS FOR OBJECTS WITH UNTAPERED ILLUMINATION AND VARYING AMOUNTS OF NOISE.

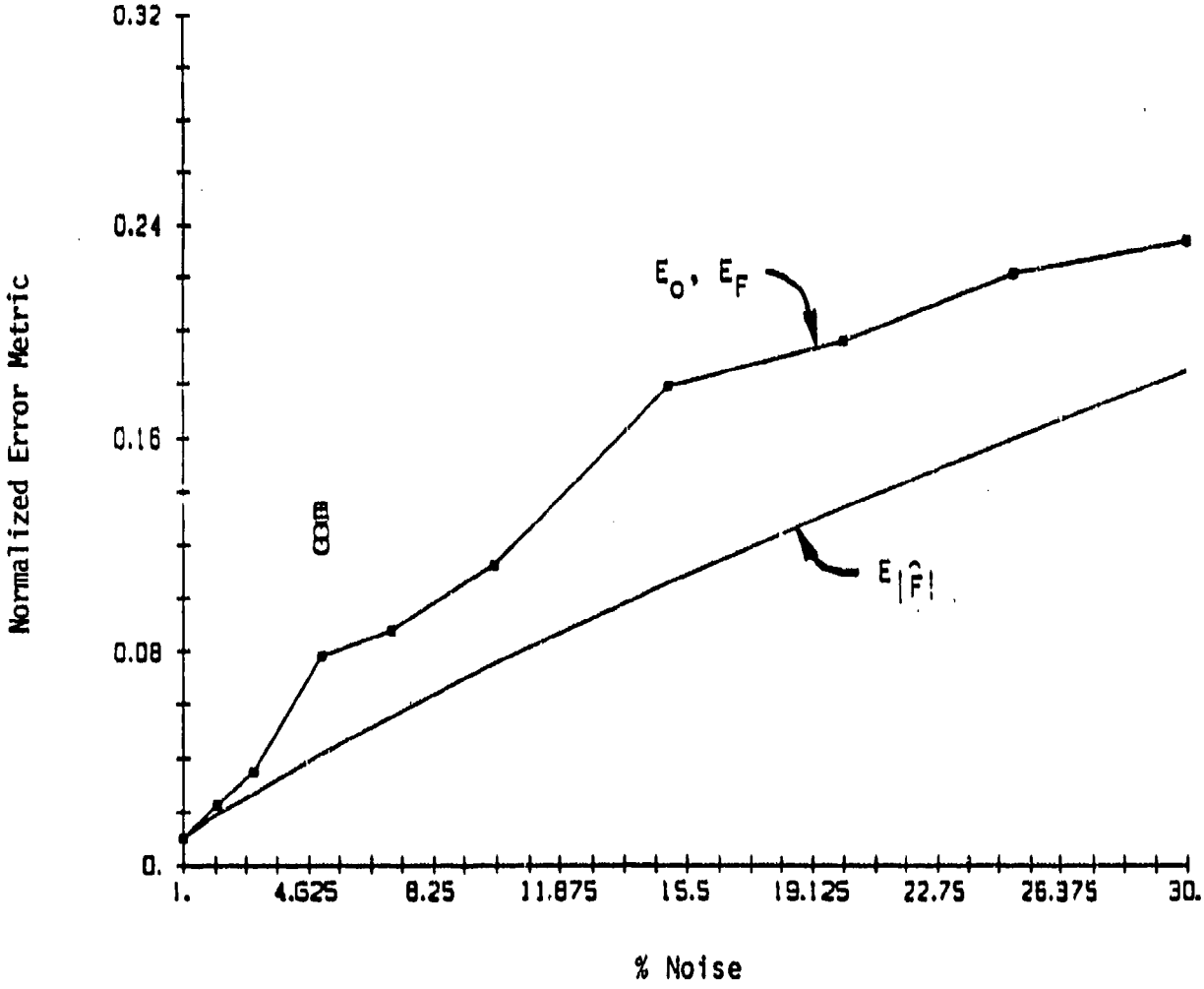


FIGURE 5-24. ALGORITHM EFFICIENCY FOR UNTAPERED ILLUMINATION.

Fourier-domain error metrics tend to be larger than the trend predicted by the other noise levels.

It is also interesting to compute a normalized absolute error in the Fourier modulus:

$$E_{|\hat{F}|} = \left[\frac{\sum_u [|\hat{F}(u)| - |F(u)|]^2}{\sum_u |F(u)|^2} \right]^{1/2} . \qquad (5\text{-}17)$$

This figure of merit serves as yet another metric for quantifying the amount of noise in the data. In addition it represents an upper bound on the achievable Fourier-domain error metric. Consider substituting the true object as the latest estimate achieved by the iterative algorithm.
If this were done, then

$$|G(u)| = |F(u)| , \qquad (5\text{-}18)$$

and it is easy to show that

$$E_F = E_{|\hat{F}|} . \qquad (5\text{-}19)$$

If we continue to iterate with error reduction we are guaranteed that E_F will not increase [5.3] and will probably decrease. The implication is that solutions exist for which the Fourier-domain error metric is at least equal to and probably less than $E_{|F|}$. Since the Fourier-domain error metrics don't achieve this upper bound, as depicted in Figure 5-24, the reconstruction quality is not yet limited solely by the noise. This comparison is a statement about algorithm efficiency, or the ability of the algorithm to achieve the upper bound. We believe that improvements to the algorithm may yet produce noise-limited reconstructions.

These experiments have examined the effect of noise in the absence of taper. They serve as a reference with which to compare experiments in which taper is present. A moderate amount of taper (6 pixels or ρ_c = 43 pixels) was selected and various amounts of noise were added to the Fourier intensity. The reconstruction sequence and the number of iterations was the same that has been used throughout the experiments with triangular illumination patterns (Table 5-1). The reconstructions with and without taper are shown in Figure 5-25. Clearly the reconstruction quality degrades much more quickly with increased noise in the presence of taper. Roughly speaking, the reconstruction with 1% noise and taper is comparable in quality to the reconstruction with 10% noise and no taper. Figure 5-26 is a plot of the absolute error in the reconstruction, E_g, as a function of the absolute error in the Fourier magnitude, $E_{|\hat{F}|}$, for both tapered and untapered objects. Evidently the inclusion of taper significantly hampers the algorithm's ability to reconstruct in the presence of noise. Figure 5-27 is a plot of E_F and $E_{|\hat{F}|}$ as a function of % noise. A careful comparison of Figures 5-27 and 5-24 shows that the algorithm efficiency, or ability to achieve the $E_{|\hat{F}|}$ upper bound, is reduced when taper is present. In other words, there is even more room for algorithmic improvement when taper is present.

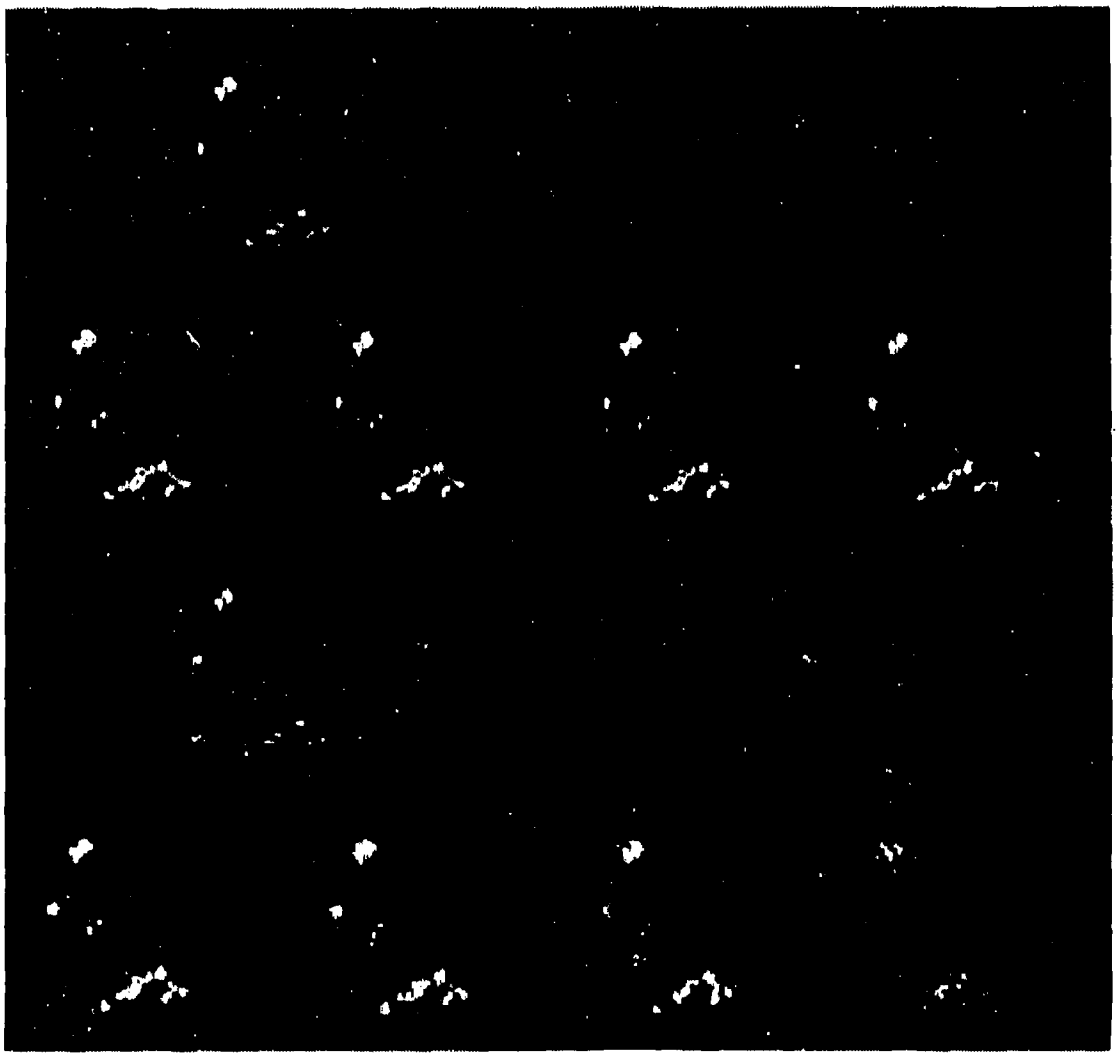


FIGURE 5-25. RECONSTRUCTIONS FROM DATA WITH VARYING AMOUNTS OF NOISE FOR UNTAPERED AND TAPERED (6 PIXELS) ILLUMINATION.

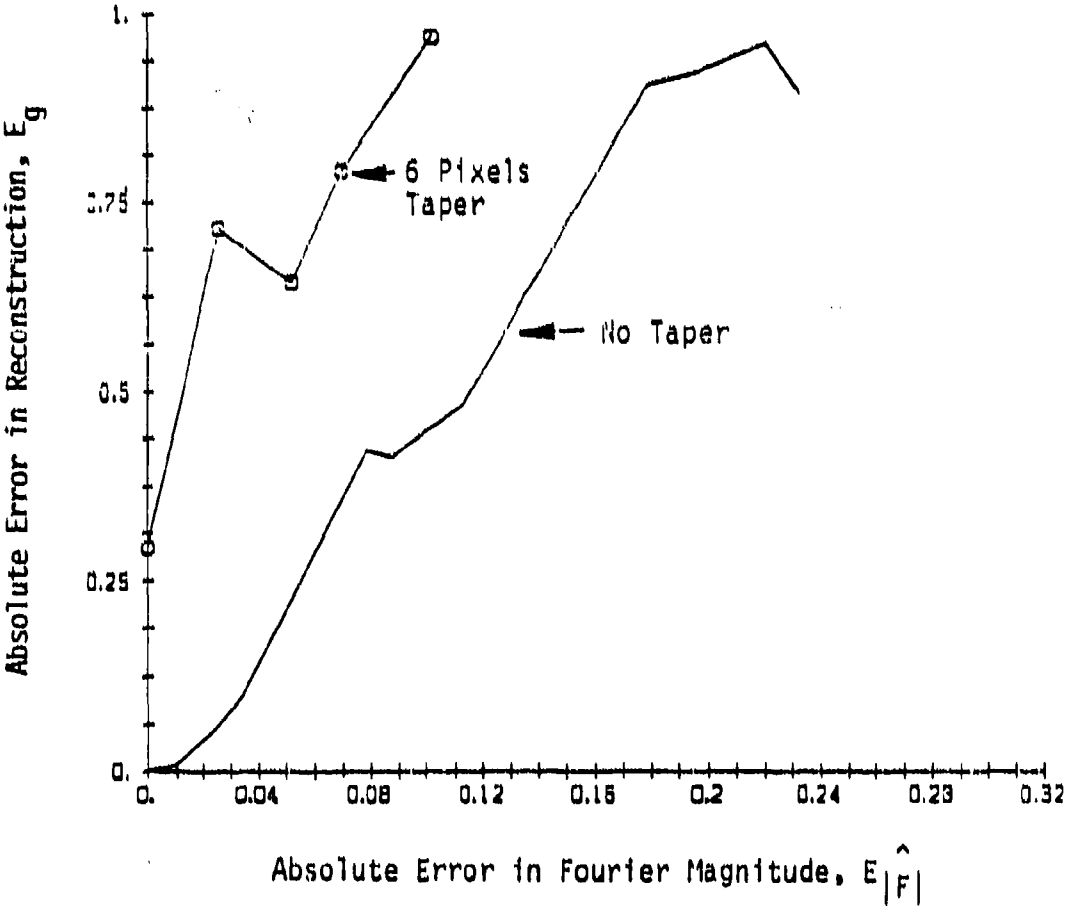


FIGURE 5-26. ABSOLUTE ERROR IN RECONSTRUCTION AS A FUNCTION OF AMOUNT OF NOISE.

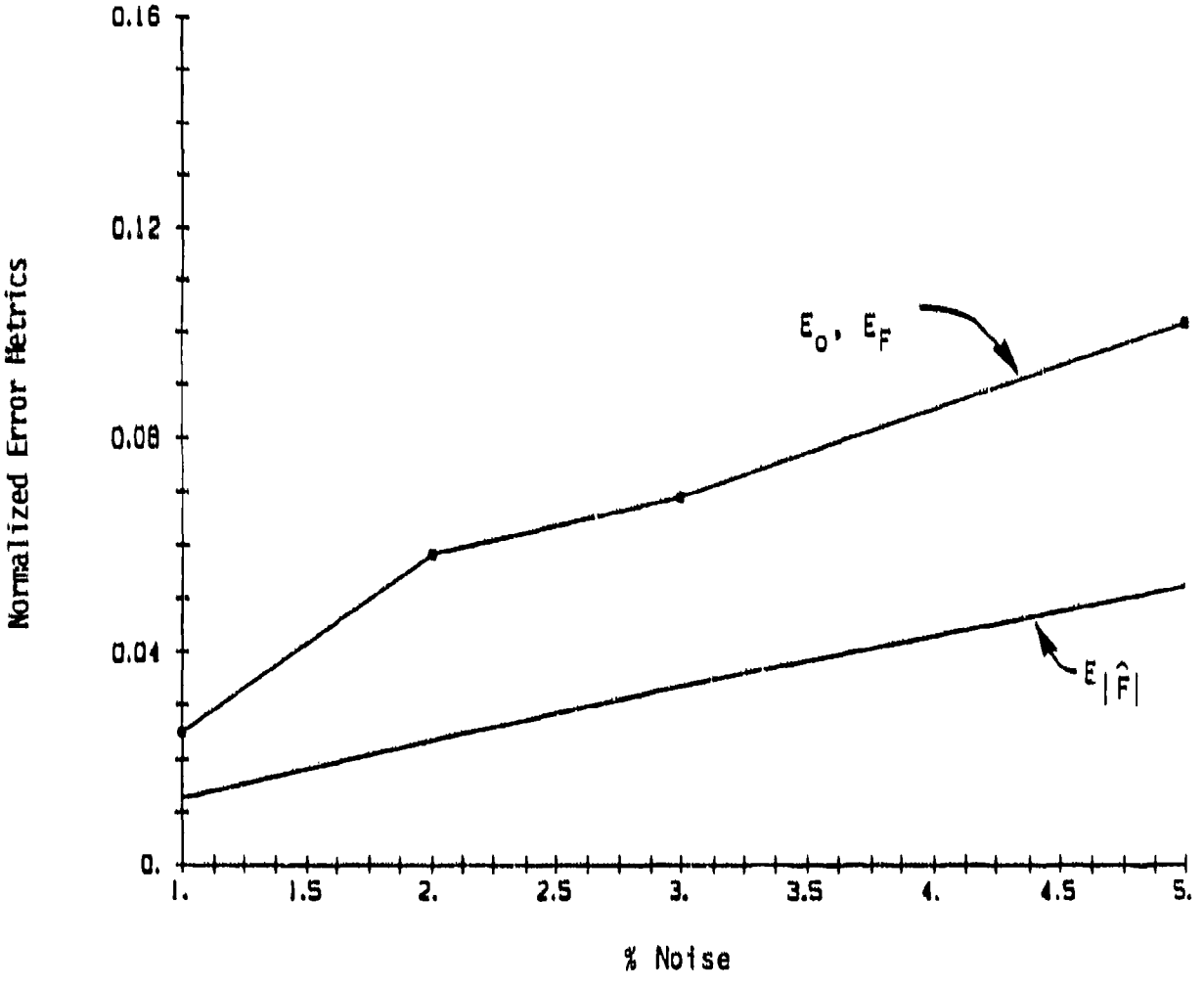


FIGURE 5-27. ALGORITHM EFFICIENCY FOR TAPERED ILLUMINATION (6 PIXELS TAPER).

REFERENCES

5.1. J.R. Fienup, "Phase Retrieval from a Single Intensity Distribution," in Optics in Modern Science and Technology, Conference Digest for ICO-13, 20-24 August 1984, Sapporo, Japan, pp. 606-609.

5.2. J.R. Fienup, "Phase Retrieval Using a Support Constraint," IEEE ASSP Workshop on Multidimensional Digital Signal Processing," Leesburg, VA, 28-30 October 1985.

5.3. J.R. Fienup, "Phase Retrieval Algorithms: A Comparison," Appl. Opt. 21, 2758-2789 (1982)

SECTION 6
GRADIENT-SEARCH METHODS IN PHASE RETRIEVAL

6.1 INTRODUCTION

Researchers have explored many approaches to solving the phase retrieval problem. These include direct methods using complex zeros in the analytically extended Fourier modulus [6.1], the error-reduction algorithm [6.2, 6.3], input-output algorithms [6.3], recursive algorithms [6.4, 6.5], and gradient-search algorithms [6.3, 6.6, 6.7, 6.8]. Of these approaches the input-output (HIO) algorithm appears to be the current algorithm of choice when operating on 2-dimensional data. The HIO algorithm has consistently outperformed competing algorithms with respect to computational burden and robustness to noise. In spite of the relative success of the input-output algorithms there are documented instances in which such an algorithm converges extremely slowly or even stagnates in its convergence [6.9].

In this report we are interested in the specific phase-retrieval problem for which the Fourier modulus and an object support constraint are known. We resurrect the idea of employing a gradient-search method in the hopes of developing an algorithm that will compete well with or complement the input-output approach. Gradient-search approaches require the determination of an object function that indicates the degree of consistency with the data and the constraints. This choice is pivotal in designing a specific gradient-search algorithm. We propose here three distinct objective functions and explore the performance of each when used in conjunction with standard gradient-search techniques. In the next section we discuss the error-reduction algorithm, the parent of the input-output algorithms and indicate how it can be interpreted as a gradient-search algorithm. We introduce the first new objective function, called the summed objective function, in Section 6.3. The second and third objective functions are introduced in Sections 6.4 and

6.5. These objective functions utilize the same object-support error metric but differ in their underlying parameters. Preliminary results with the new objective functions are described in Section 6.6. We conclude in Section 6.7 with projections of future work.

6.2 THE ERROR-REDUCTION ALGORITHM

An iterative algorithm that has enjoyed much success in phase retrieval is known as the error-reduction (ER) algorithm, which may be easily understood by referring to Figure 6-1. This algorithm consists of transforming between object and Fourier domains and applying appropriate constraints in the respective domains. We use the symbol $g_k(x)$ to represent the estimate of an object given by the kth iteration of the ER algorithm. The prime notation in $g_k'(x)$ indicates a version of the kth estimate for which the Fourier-domain constraints have been enforced. We use uppercase symbols to denote a Fourier-domain representation of a function. In practice the data are always sampled and therefore we use the discrete Fourier transform (DFT)

$$G(u) = \sum_x g(x) e^{-i2\pi u \cdot x/N} \tag{6-1}$$

and its inverse

$$g(x) = N^{-2} \sum_u G(u) e^{i2\pi u \cdot x/N}$$

in the algorithm. Of course the DFT is most efficiently computed with a fast Fourier transform (FFT). In Eqs. (6-1) and (6-2) x and u are two-dimensional vectors in the object and Fourier domains, respectively, and the summation notation is understood to represent a separate summation for each component of the vector running from 0 to N-1.

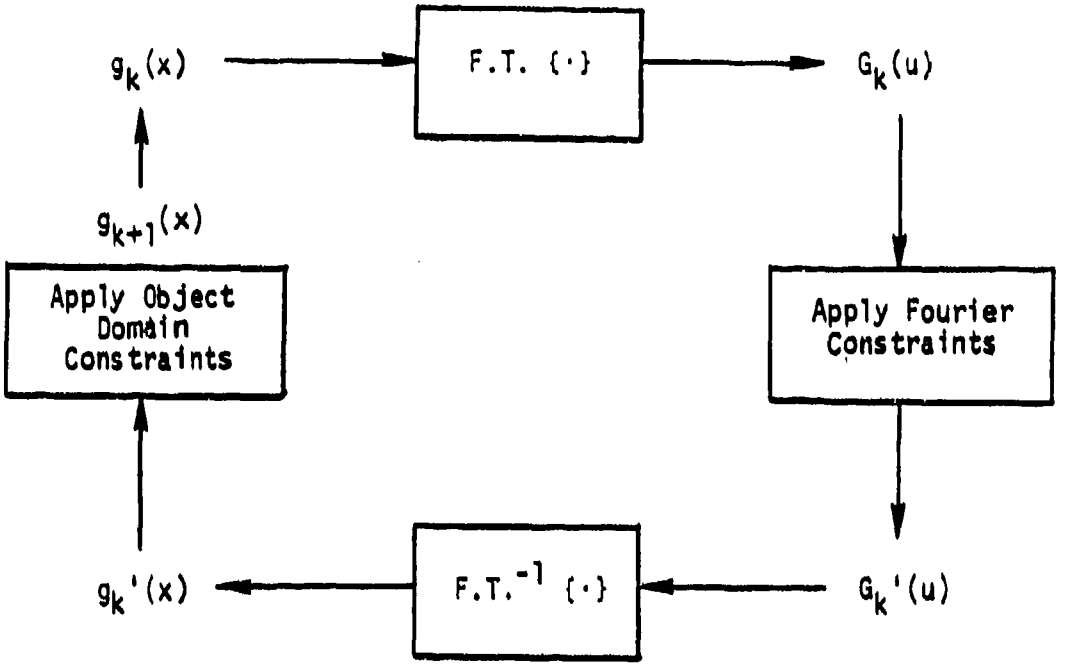


FIGURE 6-1. ERROR REDUCTION ALGORITHM

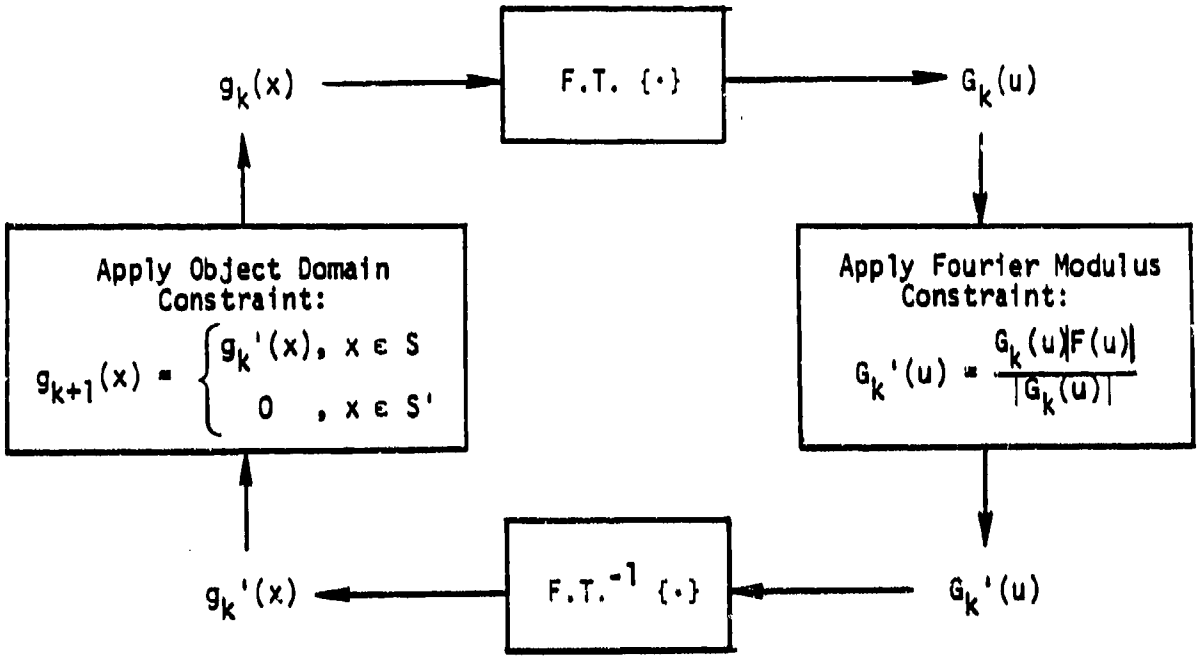


FIGURE 6-2. ERROR REDUCTION ALGORITHM FOR FOURIER MODULUS AND OBJECT SUPPORT CONSTRAINTS

In order to monitor the progress of the ER algorithm it is useful to define an error metric for each of the constraints. The error metric is essentially a mean squared error between estimates before and after a constraint has been applied and indicates the degree of agreement between the latest estimate and the known constraint. The error metric for the Fourier modulus constraint is defined as follows:

$$e_F^2 = N^{-2} \sum_u [|G(u)| - |F(u)|]^2 \quad . \tag{6-5}$$

The error metric for the object-support constraint is given by

$$e_o^2 = \sum_{x \in S'} |g'(x)|^2 \quad . \tag{6-6}$$

As the algorithm proceeds both of these error metrics will decrease. If they simultaneously achieve values close to or equal to zero then the algorithm has achieved a restoration that has good agreement with both constraints.

Suppose that we treat the error metric e_F^2 as an objective function to be used in a gradient-search algorithm. Our desire is to minimize the objective function by varying a set of parameters in the estimate. The parameters we employ are the individual pixel values of the estimate. For the present we treat only real-valued objects which require N^2 independent parameters for an NxN image (complex objects require $2N^2$ parameters). The j^{th} pixel in the object domain is located by a vector x_j where the subscript j represents any convenient ordering of the N^2 pixels. We construct an N^2-dimensional Euclidian vector space for which each coordinate axis corresponds to an individual parameter. Each point in this parameter space therefore corresponds to an object estimate and may be represented by the parameter vector g(x). We represent the j^{th} parameter and its associated parameter space unit

vector by $g(x_j)$ and v_j, respectively. The unit vector v_j may be interpreted as an estimate for which all pixels are zero except for the j^{th} pixel which has unit strength. This vector may also be represented by the Kronecker delta δ_{x,x_j}. The objective function, $e_F^2(g(x))$, is a function of the N^2 parameters, and may be visualized as a surface in an N^2+1-dimensional space. If we were able to calculate the gradient of this surface at given estimate locations then well-known gradient-search methods could be employed. The gradient is formally expressed

$$\nabla e_F^2(g(x)) = \sum_{j=1}^{N^2} \frac{\partial e_F^2}{\partial g(x_j)} v_j \quad . \tag{6-7}$$

One method of computing the gradient is to proceed numerically using a finite differences approximation to the partial derivative:

$$\frac{\partial e_F^2}{\partial g(x_j)} \approx \frac{e_F^2(g(x) + \alpha v_j) - e_F^2(g(x))}{\alpha} \tag{6-8}$$

where α is small compared with significant feature sizes in the objective surface. This brute-force approach is computationally prohibitive since each evaluation of e_F^2 involves an NxN FFT and this must be accomplished for each of the N^2 parameters. Fortunately Fienup [6.3] showed that the exact partial derivative may be calculated analytically as follows:

$$\frac{\partial e_F^2}{\partial g(x_j)} = 2\left[g(x_j) - g'(x_j)\right] \quad . \tag{6-9}$$

We reemphasize that the prime indicates that the Fourier magnitude constraint has been applied to the estimate. If Eq. (6-9) is

substituted into Eq. (6-7) the result implies that the entire gradient may be evaluated with a forward and an inverse FFT:

$$\begin{aligned} \nabla e_F^2(g(x)) &= \sum_j 2\left[g(x_j) - g'(x_j)\right]v_j \\ &= 2[g(x) - g'(x)] \end{aligned} \quad . \qquad (6\text{-}10)$$

This desirable result means that a gradient search method could realistically be employed for the $e_F^2(g(x))$ objective function.

Perhaps the simplest gradient-search algorithm is the method of steepest descent [6.10]. According to this approach the latest estimate may be improved upon by moving in parameter space in a direction opposite that of the gradient. The location of the minimum of the objective function along the resulting one-dimensional cut is then determined giving an improved estimate. This procedure is repeated iteratively until a local minimum in the objective function is achieved.

Some optimization problems afford additional a priori information about disallowed regions in parameter space. There are many ways of constraining the final solution to the allowed region of parameter space. One obvious way of incorporating this information is to proceed as usual with the steepest-descent algorithm until an estimate is produced that violates the a priori knowledge. A constraint operator is then employed to find the closest allowed estimate. The steepest-descent algorithm is then applied to the latest allowed estimate. Unfortunately this constrained steepest-descent algorithm can be very slow since the direction of steepest descent is often in competition with the direction enforced by the constraint operator.

A careful analysis of the ER algorithm reveals that it is, in fact, a constrained steepest-descent algorithm for which the objective

function is $e_F^2(g(x))$ and the knowledge of object support defines a disallowed region in parameter space. The kth iteration of the ER algorithm begins with an estimate, $g_k(x)$, and replaces its Fourier modulus with the known Fourier modulus to get $g_k'(x)$. Notice that this intermediate result is equivalent to moving from $g_k(x)$ in parameter space in a $g_k'(x) - g_k(x)$ direction; that is, in a direction opposite to that of the gradient. In fact it can be shown that the objective function is a minimum (zero) at $g_k'(x)$. Typically $g_k'(x)$ will violate the known support and therefore exists in a disallowed region in parameter space. Applying the support constraint to $g_k'(x)$ produces a new estimate, g_{k+1}, that now resides in the allowed region, thus completing one iteration of the constrained steepest-descent algorithm.

While we have thus far treated the Fourier-domain error metric as an objective function we could just as easily have selected the object-domain error metric, $e_o^2(g'(x))$, for that role. The gradient for this objective function is easily obtained because the calculation of the partial derivative with respect to a pixel value is more direct:

$$\frac{\partial e_o^2}{\partial g'(x_j)} = \frac{\partial}{\partial g'(x_j)} \sum_{x \in S'} [g'(x)]^2$$

$$= \begin{cases} 0 & x_j \in S \\ 2g'(x_j) & x_j \in S' \end{cases} . \qquad (6\text{-}11)$$

Recall that the support constraint operator sets to zero all pixels in S' and leaves those in S untouched. Clearly, this operation moves the latest estimate $g_k'(x)$ in a direction opposite that of $\nabla e_o^2(g'(x))$. In addition this objective function is quadratic along this one-dimensional cut with a minimum value (zero) at $g_{k+1}(x)$. The Fourier modulus constraint may now be interpreted as the operator that takes $g_{k+1}(x)$ out of a new disallowed region in parameter space. Thus the ER algorithm

qualifies as a constrained steepest-descent algorithm from this new perspective as well.

6.3 THE SUMMED OBJECTIVE FUNCTION

Historically the error metric e_0^2 has been used to evaluate an estimate for which the Fourier constraints have been satisfied. Consequently, this error metric is a function of the pixel values in a primed estimate, as defined in Eq. (6-6). A simple generalization of this definition yields as a new error metric that can be applied to any estimate $g(x)$:

$$\epsilon_0^2(g(x)) = \sum_{x \notin S'} [g(x)]^2 \qquad (6\text{-}12)$$

It is easy to show that the partial derivative of ϵ_0^2 with respect to pixel values in the estimate has the same form as given in Eq. (6-11). Clearly this generalized objective function and its gradient still pertain to functions for which the Fourier constraints have been satisfied. Notice, however, that $\epsilon_0^2(g(x))$ now has the same underlying parameters as $e_F^2(g(x))$. This observation affords still a third interpretation of the ER algorithm that yields new insight. The ER algorithm may be viewed as alternately performing steepest-descent operations on two objective functions, $e_F^2(g(x))$ and $\epsilon_0^2(g(x))$, that coexist in the same parameter space. In practice it is often observed that the ER algorithm converges rapidly for iterations early in the sequence but that convergence becomes painfully slow as the iteration number increases. This is because the work performed in minimizing the $e_F^2(g(x))$ objective function is largely nullified when minimizing the $\epsilon_0^2(g(x))$ objective function, and vice versa. Figure 6-3a illustrates this point pictorially. This viewpoint suggests the definition of a new objective function that is the sum of the opposing objective functions:

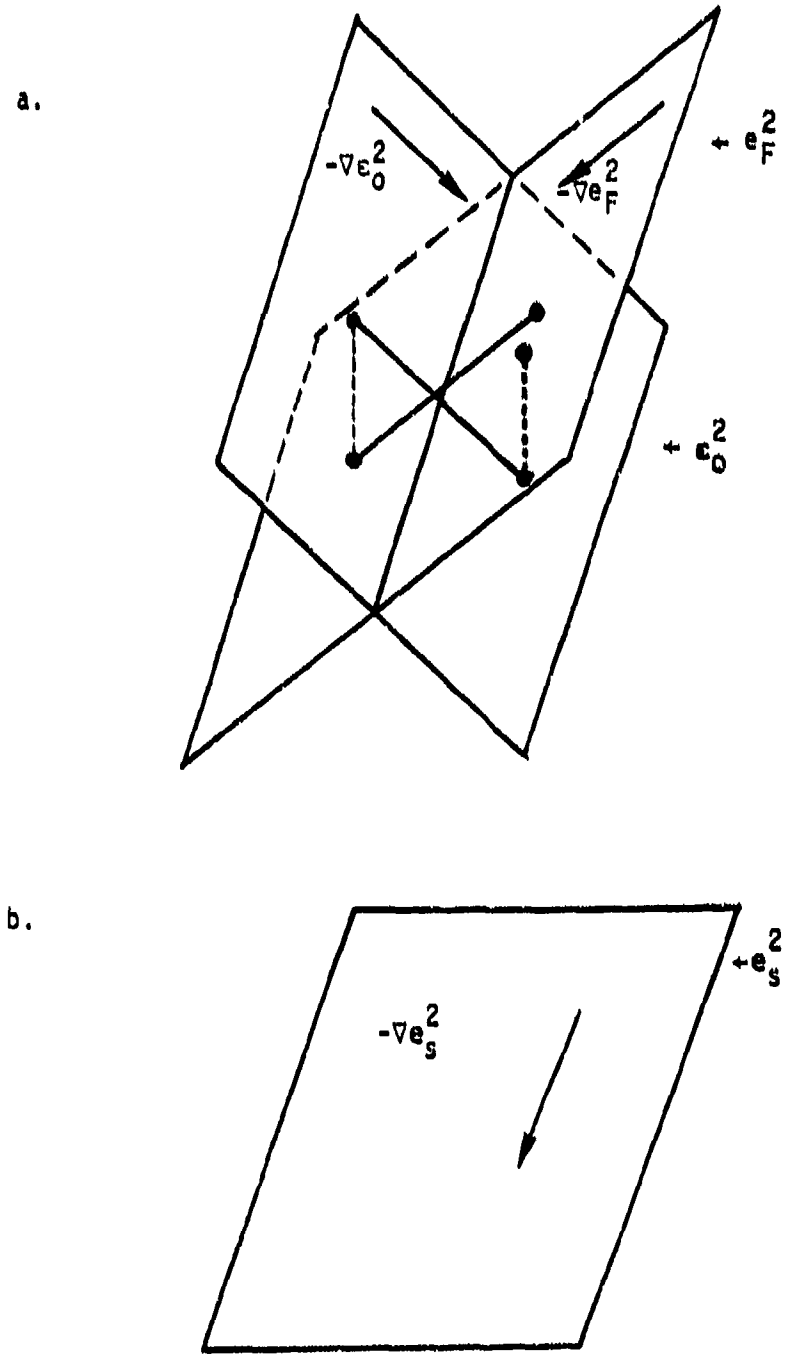


FIGURE 6-3. OBJECTIVE FUNCTION SURFACES FOR TWO PARAMETER OBJECTS a. Surfaces used in error reduction. b. Summed objective function surface.

$$e_s^2(g(x)) = e_F^2(g(x)) + e_o^2(g(x)) \quad . \tag{6-13}$$

The gradient of this new objective function is simply the sum of the gradients already derived:

$$\nabla e_s^2(g(x)) = \nabla e_F^2(g(x)) + \nabla e_o^2(g(x)) \quad . \tag{6-14}$$

The calculation of this new gradient involves a forward and inverse FFT and a small amount of computational overhead. Figure 6-3b suggests how moving in a direction opposite of the gradient of the summed objective function may circumvent stagnation due to opposing constraints. Notice that if we choose to remain with steepest descent using e_s^2, the stepsize still has to be determined. This can be accomplished by one of a variety of line search methods that utilize additional samples of the objective function. Each additional objective-function evaluation requires a single forward FFT. Furthermore, because the gradient of the summed objective function is so easily computed more sophisticated gradient-search methods such as the method of conjugate gradients or a memoryless quasi-Newton method [6.10] may profitably be employed. Finally, a simple generalization of these ideas to include complex objects is found in Appendix E.

6.4 THE $e_o^2(g(x))$ OBJECTIVE FUNCTION

We now briefly review the basic characteristics of the so-called input-output phase-retrieval algorithms. These observations will suggest the defining of a new objective function that will serve as an alternative to the summed objective function.

It is convenient to partition an iteration of the ER algorithm into two steps. The first step enforces the Fourier-domain constraints while the second step enforces the object-domain constraints. For the moment

we focus on the first step. This step involves a Fourier transformation of the latest estimate, a substitution of the Fourier modulus by the known values, and an inverse Fourier transformation. Together, these operations constitute the enforcement of Fourier knowledge and may be viewed as a single nonlinear operation. This is depicted schematically in Figure 6-4. It is important to recognize that any output of this operation will satisfy the Fourier-domain constraints and consequently e_F^2 will be zero. Should the output also satisfy the object-domain constraints then a solution has been found. This suggests that clever adjustments to the input function might produce an output that more closely satisfies the object-domain constraints. The degree of consistency with the support constraint can be monitored by the e_O^2 error metric defined in Eq. (6-6). A variety of feedback strategies borrowed from nonlinear-systems control theory can be employed to modify the latest input in order to drive the e_O^2 error metric toward zero. The use of each feedback rule defines an individual algorithm and the collection of feedback rules defines the class of input-output phase-retrieval algorithms. All feedback rules that have been employed to date are point operations meaning that an input pixel-value adjustment is based solely upon the desired change in the corresponding output pixel value.

We recognize immediately that if a solution were to serve as an input function then it will pass through the nonlinear modulus operator unchanged. Notice however that other inputs can also output a solution. In fact any input function with the proper Fourier phase will produce a solution. Thus a solution will result from any of an uncountable infinity of input functions, many of which differ dramatically from the solution. The ER algorithm may be viewed as a particular input-output algorithm for which the feedback rule drives the input (as well as the output) toward a solution. By contrast most input-output algorithms have a more flexible feedback rule since they may converge upon any of the many input functions that yield a true solution upon output.

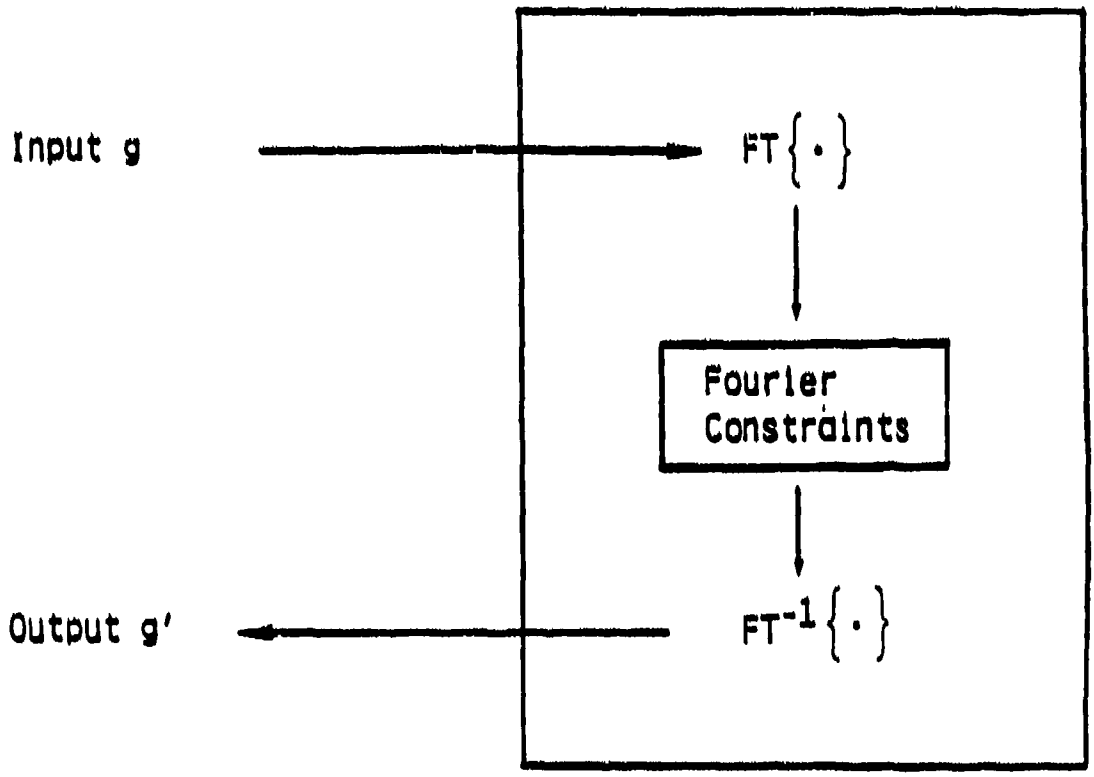


FIGURE 6-4. INPUT-OUTPUT ALGORITHM

We reiterate that any output for which e_o^2 is zero will be a solution. Therefore, the task of simultaneously minimizing the Fourier and object-domain error metrics has been converted into minimizing a single error metric. Unlike the summed objective function, however, this blending of the two error metrics into one is accomplished without resorting to ad hoc methods such as summing.

We use the term objective function to refer to an error metric in conjunction with a set of underlying parameters. A logical candidate for an alternative objective function suggested by input-output algorithms is the e_o^2 error metric as a function of <u>input</u> pixel values. This new objective function should not be confused with the $e_o^2(g'(x))$ objective function used in the ER algorithm which treats the N^2 object-estimate (output) pixel values as parameters. By contrast the new objective function, $e_o^2(g(x))$, utilizes the input-function pixel values as parameters associated with the object estimate given upon output. Having made this subtle but critical distinction we may now write an expression for the gradient of the $e_o^2(g(x))$ objective function:

$$\nabla e_o^2(g(x)) = \sum_{j=1}^{N^2} \frac{\partial e_o^2}{\partial g(x_j)} v_j \quad . \tag{6-15}$$

As before a numerical computation of the gradient is overwhelming. It is natural to ask if an analytic expression for the gradient can be derived. While the details of this calculation are outlined in Appendix F, we give the surprisingly simple result here:

$$\frac{\partial e_o^2}{\partial g(x_j)} = \sum_u \left[\frac{|F(u)|G_e(u)}{|G(u)|} - \frac{G'(u)G_e^*(u)}{G^*(u)} \right] e^{i2\pi u \cdot x_j/N} \tag{6-16}$$

where * denotes complex conjugate and $G_e(u)$ is the Fourier transform of an error image $g_e(x)$, where

$$g_e(x) = S'(x)g'(x) \tag{6-17}$$

and

$$S'(x) = \begin{cases} 1 , & x \in S' \\ 0 , & x \in S \end{cases} \tag{6-18}$$

Three FFT operations are required to compute G_e from g. A very important feature of the analytic partial derivative quoted in Eq. (6-16) is that it has the form of a DFT. The implication is that given the expression within the brackets all partial derivatives needed to compute the gradient are provided by a single DFT. Thus the total computational cost of finding $\nabla e_o^2(g(x))$ for a given input function is four FFTs plus minor overhead. With these manageable computational requirements the $e_o^2(g(x))$ objective function may be minimized via various gradient-search algorithms. Some care must be taken in the evaluation of Eq. (6-16) to avoid division to zero. This problem can be circumvented by adding a small constant to the Fourier magnitude of the input function at those spatial frequencies for which $|G(u)|$ is identically zero.

Notice that, like input-output algorithms, there are many input functions to which a gradient-search algorithm can converge for this objective function. This means that the objective function contains many global minima, each equally acceptable for producing a solution as an output. It is conceivable that this multiplicity of input solutions could yield faster convergence rates than an objective function having only a single global minimum (e.g. the summed objective function).

It is useful to recognize that any gradient-search algorithm used in conjunction with the $e_o^2(g(x))$ objective function may also be interpreted as a particular feedback rule in an input-output algorithm. Unlike other feedback rules, however, this rule is not a point operation. In other words, the gradient-search feedback rule is more flexible than other existing rules since many input pixels may be adjusted in order to effect a desired change in a single output pixel.

Unfortunately, there is no guarantee a priori that the $e_o^2(g(x))$ objective function has a surface contour that lends itself to minimization via gradient search. For example the $e_o^2(g(x))$ surface may contain many local minima in which gradient-search algorithms could become entrapped. Answers to such questions are often the by-product of extensive experimentation.

Some preliminary experiments were performed in which the $e_o^2(g(x))$ objective function was used in conjunction with the method of steepest descent. A number of observations can be made about the results displayed in Figure 6-5. Notice the dominant stripes in the gradient image for the first iteration. By gradient image we mean the image for which each pixel value is assigned the value of the associated component of the gradient. This is the image that is scaled and added to the latest input image to acquire the succeeding input image in a steepest-descent scheme. These stripes are intriguing; but their origin is unkonwn at present. The magnitude of the gradient was observed to decrease with iteration number. As a result, the stripes from the first gradient image still persist in the 100th input image. Notice, however, that the stripes do not appear in an output image, which is consistent with the notion that the input image need not resemble the output image. It is encouraging that after 100 iterations the output image bears a rough resemblance to the true object. More experimentation with this objective function is needed before a judgment can be made about its usefulness. For example, more sophisticated gradient-search methods

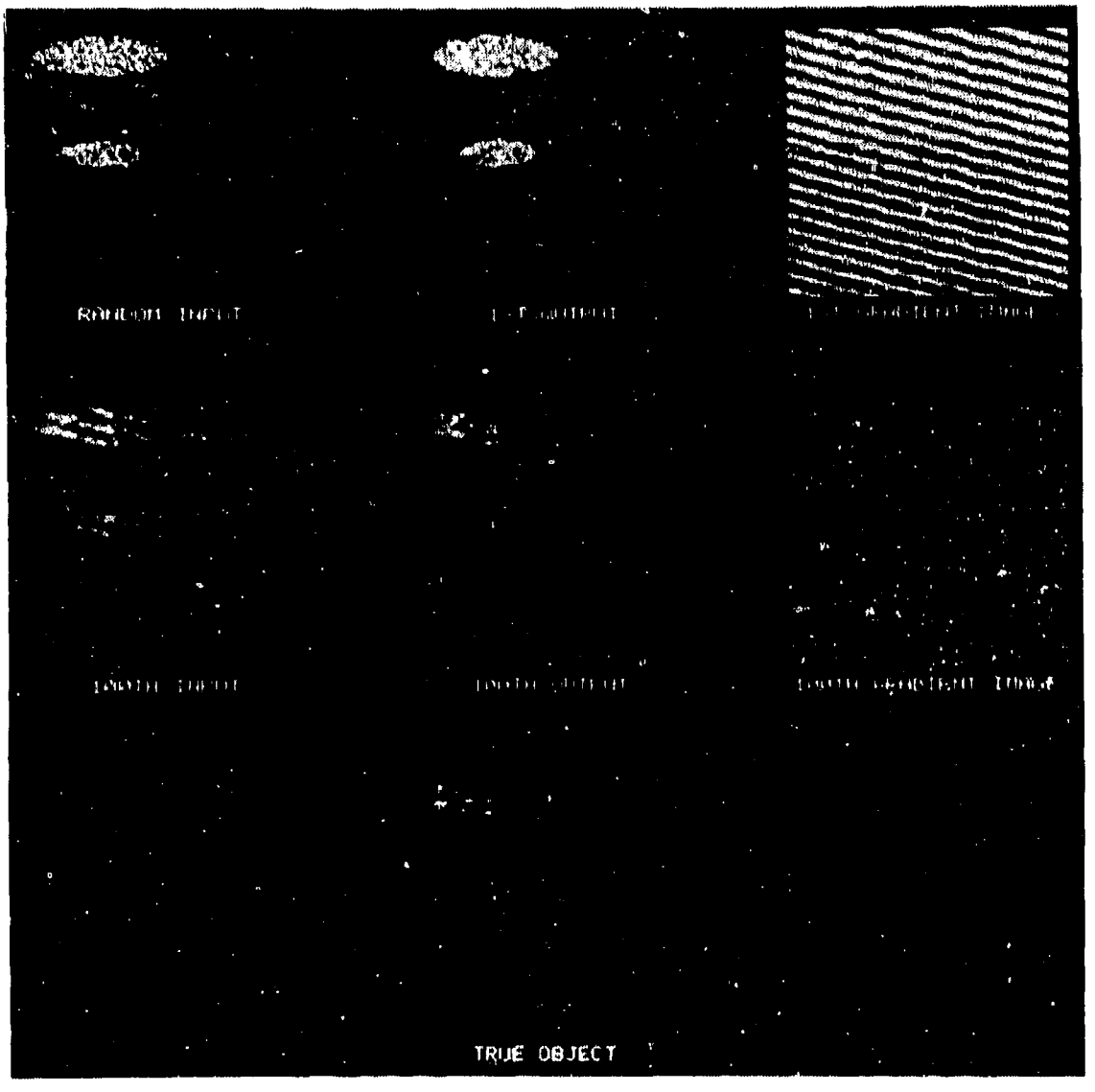


FIGURE 6-5. PRELIMINARY IMAGES DERIVED FROM MINIMIZING THE e_o^2 OBJECTIVE FUNCTION

would have a better chance of converging to a solution. Should the $e_o^2(g(x))$ objective function in conjunction with the best gradient-search methods prove not to be competitive with current input-output algorithms, it may yet be useful for breaking out of stagnation episodes.

We conclude this section by noting that while we have restricted objects to be real-valued for simplicity, the case admitting complex objects is of great interest when objects are illuminated coherently. The definition and derivation of the gradient of the $e_o^2(g(x))$ objective function for complex objects is presented in Appendix G.

6.5 FOURIER PHASE PARAMETERS

The choice of underlying parameters for an objective function can have a tremendous impact upon the behavior of gradient-search algorithms. To this point we have selected the input pixel values (or real and imaginary parts of the input pixels) as our N^2 (or $2N^2$) parameters underlying the $e_o^2(g(x))$ objective function. This choice has merit since it affords an analytic expression for the gradient requiring only four FFTs. An alternative and very different set of parameters worth consideration is the set of Fourier phase values in a Fourier estimate of a solution. Because the Fourier modulus is known, a Fourier estimate is determined by an estimate of the Fourier phase, $\phi(u)$:

$$G'(u) = |F(u)|e^{i\phi(u)} \quad . \tag{6-19}$$

An inverse FFT gives the corresponding object-domain estimate,

$$g'(x) = N^{-2} \sum_u |F(u)|e^{i\phi(u)}e^{i2\pi u \cdot x/N} \tag{6-20}$$

This estimate may also be interpreted as the output from an input-output algorithm since it has the proper Fourier modulus. Consequently, the object-domain error metric can be computed:

$$e_o^2 = \sum_{x \in S'} |g'(x)|^2 \quad . \tag{6-21}$$

The e_o^2 error metric is therefore implicitly a function of the Fourier phase values and $e_o^2(\phi(u))$ serves as the third new objective function introduced in this chapter. We mention parenthetically that throughout this section we allow for complex-valued objects since there is no simplification of derivations by resorting to real-valued objects. Notice that the designation of the Fourier phase values as the underlying parameters has fixed the number of parameters at N^2. This is exactly half the number of parameters that occur when using the real and imaginary parts of the input pixel values as parameters. It remains to be seen, though, if an analytic expression for the gradient of the object-domain error metric with respect to the Fourier phase parameters can be derived.

The gradient is defined as

$$\nabla e_o^2(\phi(u)) = \sum_{j=1}^{N^2} \frac{\partial e_o^2}{\partial \phi(u_j)} v_j \quad . \tag{6-22}$$

where v_j is the unit vector in parameter space associated with the phase parameter at location u_j in the Fourier-domain estimate. As usual, the heart of the gradient is the partial derivative

$$\frac{\partial e_o^2}{\partial \phi(u_j)} = \frac{\partial}{\partial \phi(u_j)} \sum_{x \in S'} |g'(x)|^2 \tag{6-23}$$

$$= \sum_{x \in S'} g'(x) \frac{\partial g'^*(x)}{\partial \phi(u_j)} + C.C. \qquad (6\text{-}24)$$

where C.C. stands for complex conjugate. The partial derivative in Eq. (6-24) may be simplified:

$$\frac{\partial g'^*(x)}{\partial \phi(u_j)} = N^{-2} \sum_u |F(u)| e^{-i2\pi u \cdot x/N} \frac{\partial}{\partial \phi(u_j)} e^{-i\phi(u)} \qquad (6\text{-}25)$$

$$= N^{-2} \sum_u |F(u)| e^{-i2\pi u \cdot x/N} (-i) e^{-i\phi(u)} \delta_{u,u_j} \qquad (6\text{-}26)$$

Applying the sifting property of the Kronecker delta, δ_{u,u_j}, in Eq. (6-26) leaves only one term from the summation:

$$\frac{\partial g'^*(x)}{\partial \phi(u_j)} = N^{-2} |F(u_j)| e^{-i2\pi u_j \cdot x/N} (-i) e^{-i\phi(u_j)} \cdot \qquad (6\text{-}27)$$

Substituting back into Eq. (6-24):

$$\frac{\partial e_o}{\partial \phi(u_j)} = \sum_{x \in S'} \left[g'(x)(-i) N^{-2} |F(u_j)| e^{-i2\pi u_j \cdot x/N} e^{-i\phi(u_j)} + C.C. \right] \qquad (6\text{-}28)$$

$$= N^{-2} |F(u_j)| \left[\left[(-i) e^{-i\phi(u_j)} \sum_x S'(x) g'(x) e^{-i2\pi u_j \cdot x/N} \right] + C.C. \right] . \qquad (6\text{-}29)$$

The summation in Eq. (6-29) is the Fourier error image, $G_e(u)$, defined in the previous section by Eq. (6-17). Therefore we have

$$\frac{\partial e_o^2}{\partial \phi(u_j)} = N^{-2} |F(u_j)| \left[(-i) G_e(u_j) e^{-i\phi(u_j)} + C.C. \right] \qquad (6\text{-}30)$$

$$= 2N^{-2} |F(u_j)| \, \mathrm{Im} \left\{ G_e(u_j) e^{-i\phi(u_j)} \right\} \qquad (6\text{-}31)$$

where Im {•} stands for the imaginary part.

Again we have been able to find an expression for the gradient with a remarkably compact form. Equation (6-31) implies that the component of the gradient associated with the spatial frequency u_j is proportional to the modulus at that spatial frequency and is dependent upon the Fourier-domain error image and the latest Fourier phase in a less direct way. An examination of Eq. (6-31) reveals that the entire gradient can be computed with 2 FFTs plus minor overhead. The actual evaluation of the objective function for a particular Fourier-phase estimate requires only one FFT. Thus employing Fourier-phase values as optimization parameters is certainly competitive with the use of input-pixel values from the standpoint of operations required to compute the gradient. How these two gradient-search formulations compare with respect to convergence properties can only be determined by experimentation. We might expect the Fourier-phase formulation to perform differently since the Fourier-phase parameters are so different in character from and nonlinearly related to the input pixel-value parameters. Use of the Fourier phase for parameters has the added appeal that these are in fact the unknowns in the phase-retrieval problem. As a result the Fourier phase formulation is somewhat more direct and may lend itself to analysis when noise is present.

6.6 PRELIMINARY RESULTS

A series of experiments was performed to test the use of these proposed objective functions. The first experiments performed tested the summed objective function using a conjugate-gradient minimizer. The conjugate-gradient algorithm is known to have a convergence rate far superior to that of steepest decent in the vast majority of applications [6-10]. This algorithm minimized well at the outset but after several

iterations it stagnated at an eroneous estimate. It is of considerable interest that the stagnation occurred at the same estimate for which the error reduction algorithm stagnated. Multiple experiments should be performed in an attempt to duplicate this behavior. Nevertheless this isolated result suggests that error reduction may in fact stagnate due to true local minima in the objective function as opposed to extremely slow convergence rates. Even if this conjecture were true, the summed objective function might profitably be used to replace error reduction in iterative sequences that also employ the hybrid input-output algorithm.

A second set of experiments were performed with the $e_o^2(\phi(u))$ objective function in conjunction with the conjugate-gradient minimizer. This combination managed to find the true solution for a very small object. The object support was an isosceles right triangle with seven pixels on a side. The object was embedded in a 16 x 16 pixel array. Unfortunately, when the object was doubled in size so that the base frame size was 32 x 32 pixels, the algorithm stagnated without finding the true solution. It is unclear at present why the algorithm worked for the small image and not the large image. Perhaps the additional parameters increase the probability of encountering local minima.

Preliminary results associated with the $e_o^2(g(x))$ objective function are reported in Section 6.3.

6.7 CONCLUSIONS AND FUTURE WORK

We have shown that the error reduction (ER) algorithm may be interpreted as a constrained steepest-descent algorithm for which the objective function consists of the Fourier-domain error metric as a function of pixel values in the latest estimate. In addition we have proposed three new objective functions for performing phase retrieval using gradient-search methods. These include (1) use of the summation

objective function with pixel values of the latest estimate as optimization parameters, (2) use of the object-domain error metric with input pixel values as parameters, and (3) use of the object-domain error metric with Fourier-phase values as parameters. Analytic expressions for the gradients for each of these approaches have been derived. The simplicity of these expressions implies that gradient-search methods have the hope of being computationally tractable and even competitive with existing input-output algorithms. The total number of FFTs required to evaluate the objective function and compute the gradient for each of these approaches is shown in Table 6-1.

Table 6-1
NUMBER OF FFTS REQUIRED FOR GRADIENT-SEARCH APPROACHES

Objective Function	#FFTs to evaluate objective function	#FFTs to evaluate gradient
$e_F^2(g(x))$	1	2
$e_S^2(g(x))$	1	2
$e_O^2(g(x))$	2	5
$e_O^2(\phi(u))$	1	2

Of course extensive experimentation needs to occur to see if the surface contour of each proposed objective function is well suited for gradient-search methods. Surface contour depends upon such things as the intrinsic definition of the objective function, the particular true object, and the amount of noise in the data. The suitability of a particular gradient-search algorithm to a given surface contour manifests itself in the convergence rates of the algorithm. For example a memoryless modified Newton method [6.10] may converge well with the same objective function for which a steepest-descent algorithm

stagnates. In addition, these gradient-search approaches need to be tested in a role that complements current input-output algorithms. Gradient-search approaches could make a significant contribution to the field of phase retrieval, should they consistently provide a mode of escape from any of the various types of stagnation that have been known to appear with input-output algorithms.

REFERENCES

6.1 For a list of references see Ref. 4 in D. Kohler and L. Mandel, J. Opt. Soc. Am. 63, 134 (1973).

6.2 R. W. Gerchberg and W. O. Saxton, Optik 35, 237 (1972).

6.3 J. R. Fienup, "Phase Retrieval Algorithms: A Comparison," Applied Optics 21, 2758-2769 (1982).

6.4 J. R. Fienup, "Reconstruction of Objects having Latent Reference Points," J. Opt. Soc. Am. 73, 1421-1426 (1983).

6.5 T. R. Crimmins, "Phase Retrieval for Discrete Functions with Support Constraints: Summary," OSA Topical Meeting on Signal Recovery and Synthesis II, Honolulu, Hawaii (April, 1986).

6.6 R. A. Gonsalves, "Phase Retrieval from Modulus Data," J. Opt. Soc. Am. 66, 961-964 (1976).

6.7 W. O. Saxton, Image Processing in Electron Microscopy (Academic Press, NY 1978).

6.8 R. H. Boucher, "Convergence of Algorithms for Phase Retrieval from Two Intensity Distributions," in International Optical Computing Conference, Proc. SPIE 231, 130-141 (1980).

6.9 J. R. Fienup and C. C. Wackerman, "Phase Retrieval Stagnation Problems and Solutions," Submitted for publication in J. Opt. Soc. Am. A.

6.10 D. G. Luenberger, Linear and Nonlinear Programming (Addison-Wesley, Reading, Massachusetts 1984).

SECTION 7
MODELING APPROACH TO PHASE RETRIEVAL

The modeling approach is a new method for attempting to solve the phase retrieval problem. In this section we describe the modeling approach in general terms, and then discuss a particular implementation that was attempted.

Let $F(u,v) = |F(u,v)| \exp[i\psi(u,v)]$ be the complex Fourier transform of a particular object. Suppose that either $F(u,v)$ over the entire measurement aperture or $F(u,v)$ over some small area can be modeled by a parameterized function, M:

$$M(u,v,;a,b,...) = |M(u,v;a,b...)| \exp[i\phi(u,v;a,b,...)], \quad (7\text{-}1)$$

where a,b,... are unknown parameters. If we are given only the Fourier modulus, $|F(u,v)|$, then it might be possible to estimate the phase, $\psi(u,v)$, by (1) finding the values of the parameters a,b,... that best fit the modulus of the model, $|M(u,v;a,b,...)|$, to $|F(u,v)|$, and (2) evaluating $\phi(u,v;a,b,...)$ for that set of values of the parameters.

The most difficult part of this approach is finding a model, M, that is suitable.

In a first attempt at using the modeling approach, each small area about the local maxima of the Fourier modulus was modeled using a function taken from the control theory literature. Suppose that contours about a local maximum of the Fourier modulus, at a level 3 dB down from the local maximum, have an elliptical shape, with the major axis of length w_b at an angle θ_b relative to the u-axis and minor axis w_b. Let the local maximum be at location (u_b,v_b) where it has the value

$$A_b = |F(u_b,v_b)|. \quad (7\text{-}2)$$

Also define the distance from a given point (u,v) to the peak (u_b,v_b) as

$$w = [(u-u_b)^2 + (v-v_b)^2]^{1/2}, \tag{7-3}$$

and let

$$\theta = \tan^{-1}[(v-v_b)/(u-u_b)], \tag{7-4}$$

and

$$w_c = w_b \cos(\theta_b-\theta) + w_d \sin(\theta_b-\theta). \tag{7-5}$$

Then the model we used for a region about the local maximum is

$$M(w;A_b,\theta_b,w_b,w_d,D) = \frac{A_b w_c^2}{w_c^2 - w^2 + i2ww_cD} \tag{7-6}$$

which has squared modulus

$$|M|^2 = \frac{A_b^2 w_c^4}{(w_c^2 - w^2)^2 + (2ww_cD)^2} \tag{7-7}$$

and phase

$$\phi = -\tan^{-1}[2ww_cD/(w_c^2 - w^2)]. \tag{7-8}$$

Note that the parameters w_b, w_d and θ_b are contained within w_c.

These expressions were used in the following way:

(1) A local maximum of the squared Fourier modulus was found.

(2) A curve fit of Eq.(7-7) to the squared Fourier modulus was performed to estimate the unknown parameters.

(3) The phase in that region was computed by Eq.(7-8) using the parameter estimates.

(4) Eq.(7-7) was evaluated using the parameter estimates and subtracted from the squared Fourier modulus, leaving the residual Fourier modulus.

(5) Repeat steps (1) to (4) replacing the squared Fourier modulus with the residual Fourier modulus, until all the major local maxima are accounted for.

(6) Form the net Fourier phase as the sum of all the phase functions obtained in step (3).

(7) Form an image by inverse Fourier transforming the complex function formed from the given Fourier modulus and the net Fourier phase.

Note that for large w, $|M|^2$ approaches zero and ϕ in Eq.(7-8) approaches zero, so the model has strong local effect near each local maximum and a weaker effect on neighboring points.

When the procedure was performed for a SAR image of the type used in the digital experiments described in Section 5, the reconstructed image bore no resemblance to the original object. The reason for failure is not totally understood, but we speculate that the model, Eq.(7-6), is not appropriate to the Fourier transforms of SAR images.

If further work along these lines were to be pursued, it would be important to first develop more appropriate models for SAR signal histories.

SECTION 8
LABORATORY EXPERIMENTS

Under Task 3 of the program, laboratory experiments were performed to demonstrate reduced tolerance imaging. These experiments tested the theoretical developments concerning constraints, measurements, phase retrieval and image reconstruction algorithms, and uniqueness and sensitivity issues under more realistic conditions than was possible in the computer simulations performed under Tasks 1 and 2. The use of real objects, illumination sources, optics, and detectors placed greater demands on the reconstruction algorithm. The quality of the reconstructed image from experimental data was compared to a "ground truth" image collected in the laboratory with a conventional sensor having an equivalent aperture.

Two experiments simulating different types of systems were undertaken: an active coherent experiment in the visible and a passive incoherent experiment in the visible. In the active coherent experiment, a target was illuminated with a laser and intensity data collected in the far-field of the target. A reconstruction algorithm was then used to determine the phase in the far-field of the target and therefore an image of the target. Thus, this experiment simulated an active, coherent reduced-tolerance sensor which measures intensities only and would be insensitive to aberrations of the primary optics. In the passive incoherent experiment, a multiple mirror telescope was simulated and degraded/defocused image intensities were measured. Further discussion of both experiments is given in Sections 8.1 and 8.2.

8.1 ACTIVE EXPERIMENT

The active experiment demonstrated imaging of a coherently illuminated target from intensity-only measurements made in the far-field. This simulated a sensor having greatly reduced tolerance to the position and quality of its receiving aperture compared to a

conventional imaging sensor. The wide range of parameters which were considered in planning this experiment and which were available to be varied to test theoretical developments and computer simulation results are discussed in Section 8.1.1. The experiment design and data collection are explained in Section 8.1.2. Data processing methods and experimental results are given in Section 8.1.3.

8.1.1. ACTIVE EXPERIMENT PARAMETERS

A permanent phase retrieval laboratory was created. In this laboratory, a wide range of parameters is available to be varied to test theoretical developments and computer simulation results. These parameters had to be carefully controlled to ensure meaningful results in the active experiment. The most important of these parameters are discussed in this section.

The pattern of illumination on the target can be described by its spatial and temporal coherence, shape, sharpness of edges, phase distribution, angle of incidence on the target, and polarization. All of these parameters may affect the quality of the reconstructed image. Equally important, they may take on different values depending on the application being simulated in the laboratory. In an application where the target is actively illuminated by a laser, the spatial (transverse) and temporal (longitudinal) coherence lengths may be less than the target size. The laboratory system can allow illumination with variable coherence lengths either by manipulating the spatial coherence of a gas laser or by using a broader-band dye laser. It is known from ERIM investigations that the illumination pattern shape and sharpness of the edges affects reconstruction algorithm convergence. In applications, the range of illumination shapes and sharpness of edges may be limited by practical considerations on the transmitting aperture. The laboratory system can accept a variety of masks and image them onto the target through a controllable finite aperture to control illumination

shape and sharpness. Since the illumination phase may not be constant over the target in practical applications, the masks can be holographic if experimental control of the illumination pattern phase distribution is desired. Applications may be either monostatic or bistatic, so the laboratory system allows for either. The target will, in most cases, partially depolarize the illumination. The experimental system is capable of making measurements of the two orthogonal components of the light at the detector.

The target parameters include reflectivity contrast and structure, surface roughness, surface topography (3-D nature of target), motion during the measurement process, and noncoherent background illumination. Practical targets will vary in their roughness, although nearly all will be rough at visible and infrared wavelengths. An experimental system should primarily use rough targets to create real speckle effects. However, it was useful in the active experiment to use smooth targets (film transparencies in a liquid gate) in setting up the experiment to test and debug the optical and electronic components and software. Real targets are three-dimensional, but to varying degrees. A variety of 3-D objects is available to the experimenter. Any real target will also be noncoherently illuminated from various thermal sources. While this illumination will only add a uniform bias to the far-field measurements, it can be included in the experimental setup.

The propagation path between the target and the sensor can be of such a length as to be either near or far field and can include atmospheric turbulence, scattering (aerosols, fog, smoke), and absorption. In an application, any or all of these effects may be present. The optical setup allows for insertion or removal of lenses to give either near or far field conditions at the detector. (It must be noted that the size of the speckles in the measurement plane will depend on the sensor distance.) Turbulence with the proper statistical properties is very difficult to simulate in an indoor laboratory. The

best approach (and one which allows reproducible turbulence) is to use movable phase plates. These are glass plates with controlled thickness variations. Scattering and absorption are easier to simulate (e.g., with fog chambers, optical narrow-band filters). Since their main effects are to reduce signal levels and increase detector bias light levels, they can be studied as described in the next paragraph on detector parameters.

The most important detector parameters are signal level, type of noise, noise level, background illumination level, spatial and temporal sampling rates, polarization detected, nonlinearities in response, and spatial nonuniformities in response, bias and noise. The values of these parameters are crucial to the viability of any real application. The experimental setup is able to vary the signal and background illumination levels, simulate various sampling rates, detect orthogonal polarizations, and create nonuniform background illumination levels. The types of noise, noise level, nonuniformities in detector response and noise (pattern noise), and nonlinearities in response are primarily determined by the detector chosen for the experiment. An important aspect of the experiment design and procedure was to measure, calibrate, and correct for the effects of these parameters on the measured data. This was a difficult task and much practical experience was gained which can be applied to other detectors in the future.

Many applications involving active illumination can be expected to have low signal levels. To adequately simulate these applications, an image intensifier can be used before the detector. Even with a thermal-noise-limited detector, the use of an image intensifier may allow operation in a shot (photon) noise limited mode. The image intensifier will, of course, also have nonuniformities, nonlinearities, and a spatial sampling rate which must be measured and considered in the experimental setup and data analysis.

The effects of speckle are so important to an active coherent experiment that its parameters deserve to be discussed separately. The speckle in the measurement plane will have its size determined chiefly by the size of the illuminated region on the target and by the optics (if any) placed between the target and the sensor. Ideally, the detector must sample the intensity speckle pattern at the Nyquist rate (two samples per speckle) or greater. For a given detector, magnifying optics may be necessary. It may also be desired to investigate the effect of measurements at less than the Nyquist rate.

The data processing hardware and software available in the phase retrieval laboratory is discussed in Appendix H.

8.1.2. ACTIVE EXPERIMENT DESIGN AND DATA COLLECTION

The optical system used to obtain the results shown in Section 8.1.3 is shown in Fig. 8-1. An argon-ion laser beam of wavelength $\lambda = 0.5145$ μm is spatially filtered, collimated, and used to illuminate a transmissive object (some experiments with reflective objects were also performed). The object consists of a binary mask placed in contact with a ground glass [see Fig. 8-2]; thus the transmittance of the object is binary in intensity and random in phase. An a priori image support constraint is introduced by giving the transmissive region of the mask a known overall triangular shape. The lens L_1 of focal length f_1 produces the Fourier transform of the complex-valued object transmittance in its back focal plane. There, an aperture A selects a portion of the Fourier transform and lenses L_2 and L_3 (of focal lengths f_2 and f_3, respectively) image this portion, with suitable magnification, to the detector for collection of Fourier intensity data. When the removable mirrors M_1 and M_2 are in place, the light is diverted through lens L_4 which produces an image of the object at the detector. Because of the placement of aperture A, this conventional image provides a reference

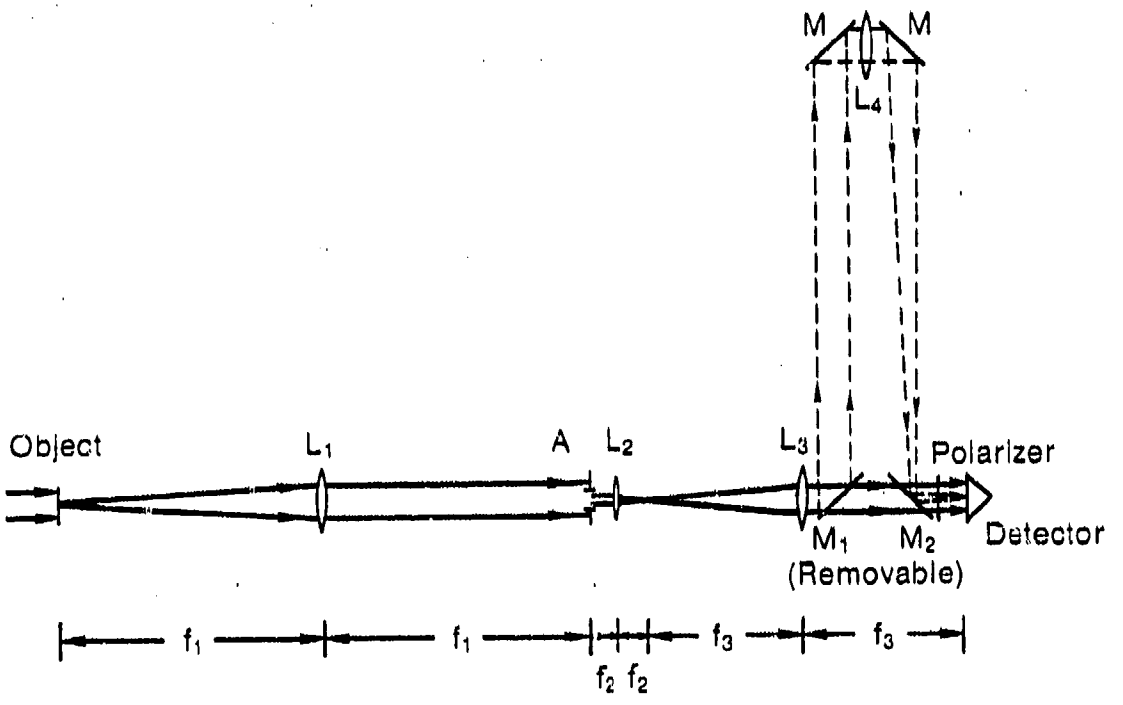


FIGURE 8-1. EXPERIMENTAL OPTICAL SYSTEM FOR ACTIVE EXPERIMENT.

FIGURE 8-2. INCOHERENT IMAGE OF TEST OBJECT.

for comparison to the image reconstructed by phase retrieval. A polarizer was placed just before the detector to ensure detection of only a single polarization.

The base and height, d, of the triangular mask and the lens focal lengths must be chosen so that the speckle in the Fourier intensity at the detector is adequately sampled. Assuming a speckle size at the detector equal to $\lambda f_1 f_3 / f_2 d$, then, to sample the intensity at the Nyquist rate, the detector pixel spacing, Δs, must equal $\lambda f_1 f_3 / 2 f_2 d$. In the experiment, d was about 16 mm and f_1, f_2, and f_3 were 500 mm, 50 mm, and 300 mm, respectively, giving $\Delta s \cong 48\ \mu m$. The CCD detector used has horizontal and vertical pixel spacings of 30 μm and 18μm, respectively, so the data were sampled at greater than the Nyquist rate.

Since the Fourier intensity has twice the spatial frequency bandwidth of the complex-valued Fourier transform, measurement of $2N_1$ by $2N_2$ Nyquist-spaced samples of the Fourier intensity enables reconstruction of a complex-valued image of N_1 by N_2 resolution elements. The number of horizontal and vertical pixels in the detector array thereby sets an upper limit on the size, in resolution elements, of the reconstructed image. In this experiment, the central 180 by 256 pixels (over a 5.4 by 4.6 mm region) of the detector were used. The aperture A was therefore 0.9 by 0.77 mm and the reconstructed and conventional images had 56 by 48 resolution elements of size 0.28 by 0.34 mm (at the object).

The diameter, D, of lens L_1 was 50 mm which exceeds the diameter $\sqrt{2}(d + 2f_1 \tan \theta) \cong 25$ mm (where $\sin \theta = \lambda N / 2d$ and N is either N_1 or N_2) needed to avoid vignetting. The other lenses need also be only f/20. The focal length, f_4, of lens L_4 was 1000 mm, which resulted in a demagnification of the conventional image by a factor of three while still allowing more than adequate sampling by the detector.

The data was detected by a Fairchild CCD 3000F television camera with a fiber optic faceplate. The RS170 video signal was converted to a 512 by 512, 8-bit digital image using an Imaging Technology IP-512 video processor. For both Fourier intensity and image data, a single video frame of data was digitized and a second dark frame was digitized and subtracted to remove pattern noise. The automatic gain control of the camera was disabled and the laser output was adjusted so that the brightest speckle nearly saturated the detector. Since the digitizer sampling rate was not matched to the detector horizontal pixel spacing, a Matthey MLW 401B low-pass video filter with a 3-dB width of 4.3 MHz was used to reduce the effect of CCD clock noise on the digitized data.

The support constraint was measured by increasing the size of the aperture A, digitizing the resulting high resolution image of the object, and measuring the base and height of the triangle in the digital image. Since the focal lengths of the lenses and the digitizer pixel horizontal spacing are not known exactly, calibration measurements must also be made to determine the spatial scaling of both the Fourier intensity and image data. This was done using an object consisting of two circular apertures separated by about four times their diameter. By orienting this object both horizontally and vertically, gathering both Fourier intensity data and high resolution image data, and using the known digitizer pixel vertical spacing (which is equal to the detector pixel spacing of 18 μm), the spatial scaling of the data may be computed. For reference, the digitizer pixel horizontal spacing was determined to be 21.2 $\pm$ 0.1 μm, yielding 256 samples over a 5.4-mm width of the detector.

To estimate the signal-to-noise ratio of the data, the detector was uniformly illuminated with an extended noncoherent source, 10 frames were digitized, and the standard deviation of the digitized values was computed on a pixel by pixel basis. The standard deviations of the 8-bit data ranged from 0.8 to 1.3. (Part of this variation could be due

to variations in light source intensity during the collection of the 10 frames.) For signals which nearly saturate the detector, the signal-to-detector noise ratio of the digital data is therefore about 200 to 1. Data for correction of spatial variations in detector response were collected by uniformly illuminating the detector at one-tenth the saturation light level, summing 10 digitized frames, and subtracting 10 dark frames. The ratio of the standard deviation to the spatial mean of the response was about 3%.

8.1.3 DATA PROCESSING AND EXPERIMENTAL RESULTS

Data processing began with a 256 by 256 array of digitized Fourier intensity data shown in Fig. 8-3. This array was divided by the response data to correct for spatial variations in detector response. The speckle contrast was then measured and found to be 80%. Since the speckle should have 100% contrast, it was assumed that some positive bias must be present in the data, possibly due to the effect of the low pass video filter. Therefore, a constant was subtracted from the data to increase the contrast to 90%. Data values which become negative in this process were set to zero. The Fourier transform of this data is the autocorrelation of the image. Since the image support is known, the autocorrelation support is also known. Therefore, to reduce high frequency noise in the Fourier intensity data, it was low-pass filtered by zeroing out those parts of its Fourier transform which lay outside the support of the autocorrelation of the known triangular image support. After setting any negative values to zero, this filtered Fourier intensity data was then square-rooted to give Fourier modulus data.

The Fourier modulus data and the triangular image support constraint were used in the iterative Fourier transform phase retrieval algorithm [8.1-8.4]. Several cycles (each of 30 to 40 iterations of hybrid input-output with β = 0.7 followed by 10 to 20 error-reduction iterations)

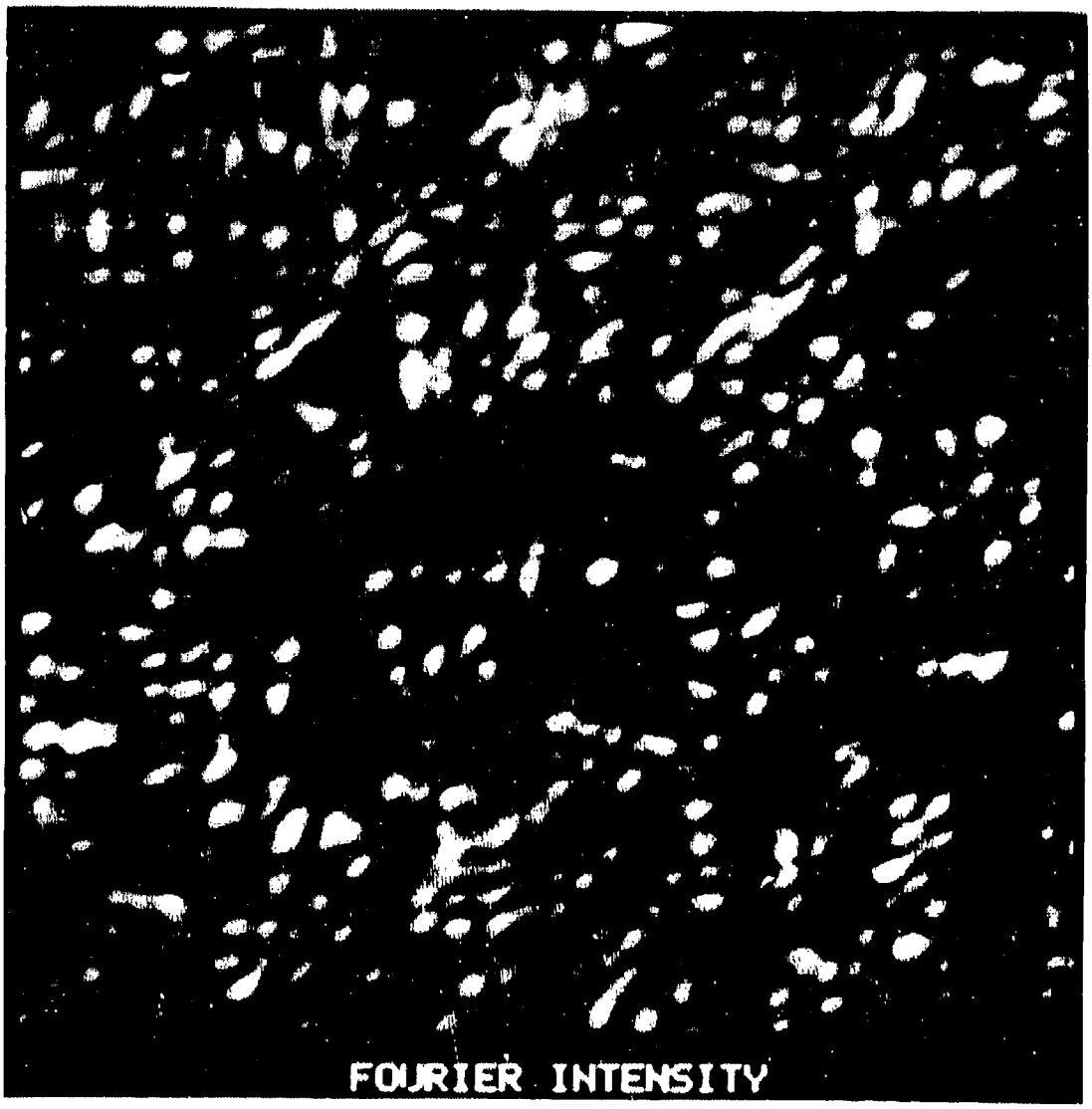


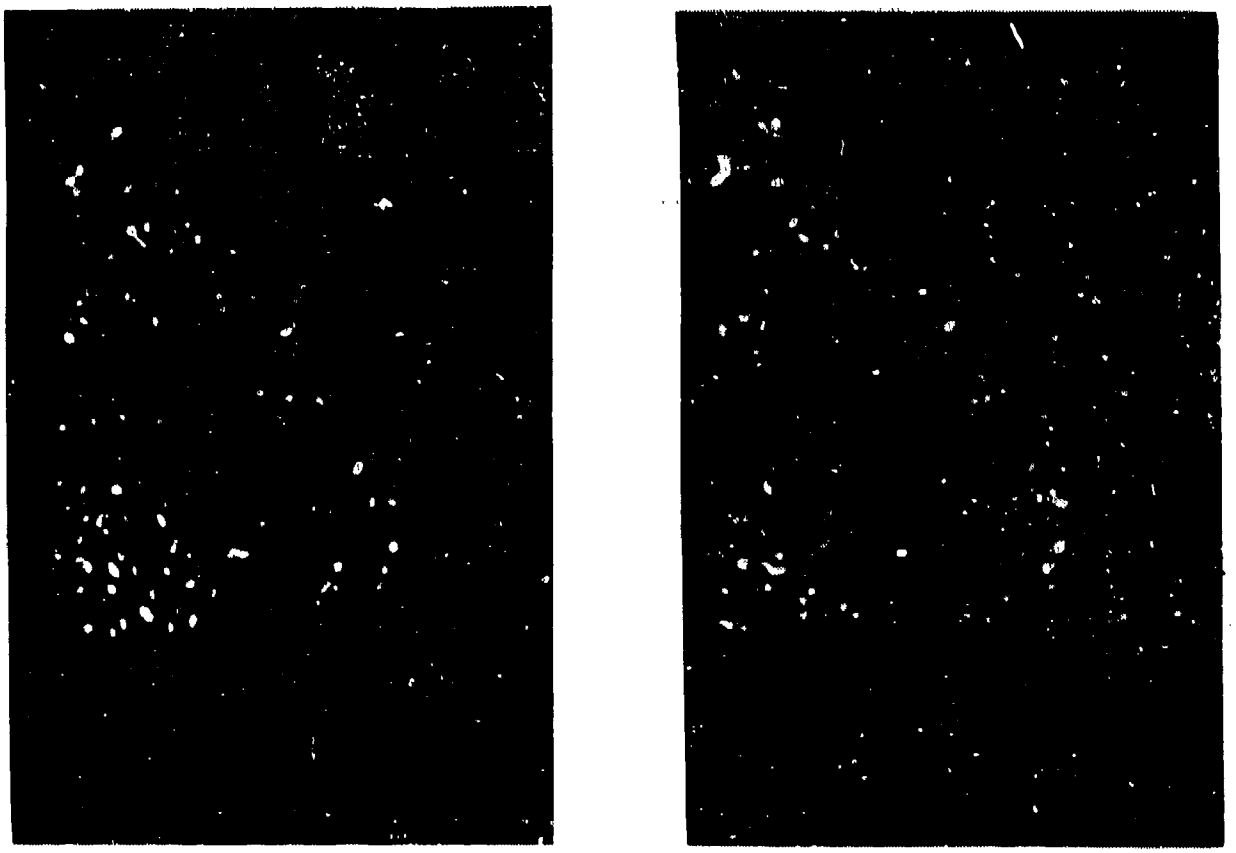
FIGURE 8-3. EXPERIMENTAL COHERENT FOURIER INTENSITY DATA.

were performed for a total of 840 iterations. At this point the algorithm was making no further progress. The normalized root-mean-squared image-domain error, E_o [8.1-8.4], was 0.112; that is, the reconstructed image was within 11% of agreeing with the measured Fourier modulus data and the support constraint. This is an indication of the amount of noise and distortion present in the measured data.

The squared-magnitude of the complex-valued reconstructed image is shown in Fig. 8-4(a) and a conventional (ground truth) image having the same spatial frequency bandwidth is shown in Fig. 8-4(b). The reconstructed image was resampled to approximately match the dimensions of the conventional image. Many of the bright speckles appear in the same location in each image indicating that some of the high frequency information, as well as the low frequency information, has been successfully reconstructed. Since several minutes elapsed between collecting the Fourier intensity data and collecting the conventional image, some of the difference between the two images may be attributable to speckle boiling during the time interval.

8.2 PASSIVE EXPERIMENT

The objective of the passive experiment is to demonstrate imaging (in the visible) of a noncoherently illuminated target from intensity-only measurements. In conventional passive imaging systems where the image is degraded by phase aberrations due either to atmospheric turbulence or to misalignment of segmented optical element arrays, it is known that image quality can be improved by using iterative reconstruction algorithms operating on <u>two</u> 2-D intensity measurements [8.5-8.6]. These two measurements can be of a best focus image and an intentionally slightly defocused one (that is, a quadratic phase error is intentionally introduced -- phase diversity). The use of this approach allows reduced tolerance to atmospheric phase or to accurate positioning and alignment of segmented optics.



(a) (b)

FIGURE 8-4. IMAGE RECONSTRUCTION FROM EXPERIMENTAL DATA. (a) Intensity of coherent image reconstructed by phase retrieval. (b) Intensity of conventional coherent image.

The passive experiment simulated an imaging system in which the image is degraded by misalignment of segmented optics in the pupil plane. Two image intensity measurements were made: one in the plane of best focus of the degraded image and one in a plane where the image is defocused. The separation of these two planes (and therefore the amount of quadratic phase error in the pupil plane giving the defocused image) was measured. An iterative algorithm (using gradient-search techniques) was used in an attempt to estimate the (phase) misalignment of the segmented optics.

The optical setup used in the experiment is shown in Fig. 8-5. The two lens imaging configuration of L_2 and L_3 allowed access to the pupil plane. A computer-generated holographic element P in the pupil plane allowed phase misalignment of segmented optics to be simulated. The six segment hologram aperture is shown in Fig. 8-6. Phase shifts of the hologram carrier frequency simulated piston errors in the segmented optics alignment. Since the hologram is computer-generated, the piston errors are known. Another advantage of the setup is that a "ground truth" image can be formed using the undiffracted light.

In order to make the laser light spatially incoherent, the unexpanded laser beam first illuminates a stationary ground glass, G_1, and then a rotating ground glass, G_2. The center of the laser beam was approximately 60mm from the axis of rotation of the ground glass. The target T used in the experiment was a segment of a standard bar target. The lens L_1 of 100mm focal length images the bar target onto an adjustable rectangular aperture A that blocks everything except elements 4 (45.3 lp/mm) and 5 (50.8 lp/mm) of group 5. Note that the 100mm focal length lens magnifies the target by roughly 2x. The two lenses L_2 and L_3 of focal lengths 50mm and 300mm image the target onto the sensor with 6x magnification. The carrier frequency of the hologram was chosen so that the undiffracted and diffracted images are separated at the detector (a Fairchild CCD3000 camera).

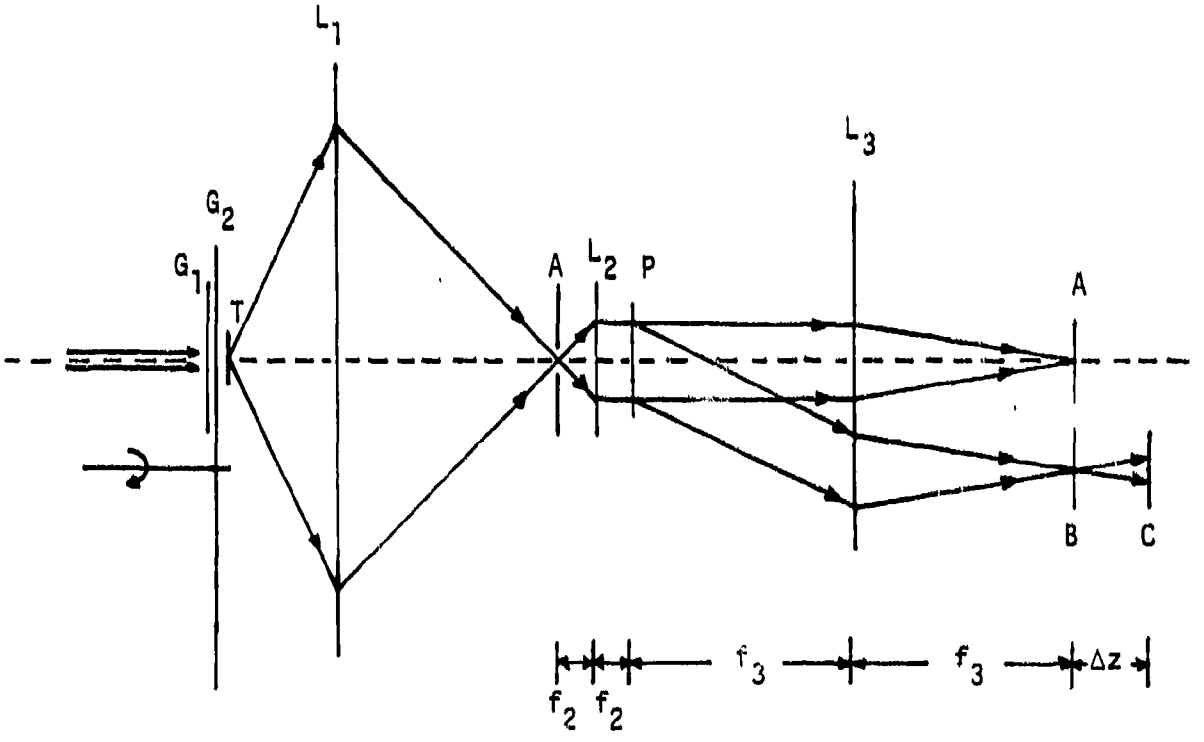


FIGURE 8-5. EXPERIMENTAL OPTICAL SYSTEM FOR PASSIVE EXPERIMENT.

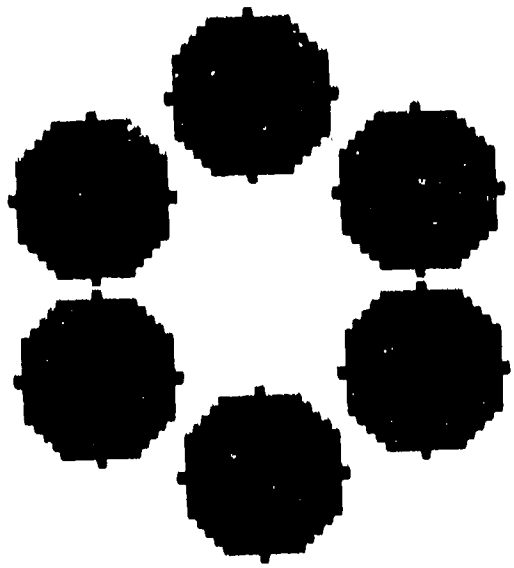


FIGURE 8-6. SIX SEGMENT HOLOGRAM APERTURE USED IN PASSIVE EXPERIMENT.

A high-resolution image of the object was collected by removing the hologram and imaging through a circular aperture of approximately 2cm diameter. The hologram was then inserted and the undiffracted image (A in Fig. 8-5) was collected. Because of the low diffraction efficiency of the hologram, all diffracted images were obtained by adding 32 frames of data, then blocking the laser beam and subtracting 32 dark frames. A best focus image (B) was digitized and images (C) at 1, 2, and 3 waves of defocus also collected. To compute the defocus distance Δz, the following formula was used,

$$\Delta z = \frac{1}{1/f - 2n\lambda/\rho^2} - f, \qquad (8\text{-}1)$$

where f is the focal length of the second imaging lens (300 mm), n is the number of waves defocus, ρ is the radius of the hologram (1.22 mm) and $\lambda = 0.5145\ \mu m$.

The final data set collected was a responsivity map of the detector array. A distant incoherent source illuminated the array at a light level close to the peak value for the diffracted images. The top and right sides of the array were masked to provide reference regions. Thirty-two frames were added and 32 dark frames were subtracted to give the responsivity map. This data was processed using an iterative gradient search algorithm in an attempt to estimate the (known) piston errors. Although several variations in the choice of defocus data and in the initial starting point for the search have been tried, to date no successful (low error) estimate has resulted.

8.3 CONCLUSIONS

Phase retrieval from Fourier intensity data from a coherently illuminated, diffuse object collected in a laboratory experiment has been demonstrated. The image-domain constraint for the iterative

Fourier transform reconstruction algorithm was a known triangular support. The reconstructed image shows good agreement with a conventionally obtained image. With this promising beginning, further laboratory experiments are warranted. Image reconstruction accuracy should be investigated as a function of, for example, object support shape, sharpness of edges of support, object contrast, object surface roughness, detected data signal-to-noise ratio, data sampling rate, detector calibration, data filtering, and iterative algorithm versions used.

Laboratory data was collected for an incoherently illuminated object imaged through a holographically-simulated misaligned segmented optical system. A lack of sufficient funds prevented us from refining the experiment to the point where useful imagery could be reconstructed from the laboratory data, and so no conclusions can be drawn from that experiment. Nevertheless, we remain optimistic that with further refinements the technique of phase diversity can allow us to recover diffraction-limited images from badly misaligned segmented-aperture telescopes, as had been demonstrated earlier by computer simulations. We recommend that further experiments be carried out to verify the usefulness of this concept. If successful, this concept could dramatically increase the performance or decrease the cost of segmented optical imaging systems.

REFERENCES

8.1 J.R. Fienup, "Reconstruction of an Object from the Modulus of Its Fourier Transform," Opt. Lett. 3, 27-29 (1978).

8.2 J.R. Fienup, "Phase Retrieval Algorithms: a Comparison," Appl. Opt. 21, 2758-2769 (1982).

8.3 J.R Fienup and C.C. Wackerman, "Phase Retrieval Stagnation Problems and Solutions," J. Opt. Soc. Am. A 3, 1897-1907 (1986).

8.4 J.R. Fienup, "Reconstruction of a Complex-Valued Object from the Modulus of Its Fourier Transform Using a Support Constraint," J. Opt. Soc. Am. A 4, 118-123 (1987).

8.5 R.A. Gonsalves, "Phase Retrieval and Diversity in Adaptive Optics," Opt. Eng. 21, 829-832 (1982).

8.6 R.G. Paxman and J.R. Fienup, "Image Reconstruction for Misaligned Optics Using Phase Diversity," J. Opt. Soc. Am. A 3, P5 (1986).

SECTION 9
APPLICATION AND ARCHITECTURE STUDY

9.1 INTRODUCTION

In this section, the practicality of the reduced tolerance imaging concept is investigated for a tactical imaging application using laser illumination. The optical and data processing architectures are defined for the imaging system and relationships are determined between system performance measures and system design parameters. For a specific wavelength and desired performance, numerical values of the system design parameters are computed.

9.2 TACTICAL IMAGING APPLICATION

The application to be considered is that of tactical imaging using laser illumination. As shown in Fig. 9-1, both air-to-ground and ground-to-air imaging can be treated by the same analysis although the primary application is air-to-ground since it is required that the area of reflectivity fill the illumination pattern. A pulsed laser would be used to illuminate the target region and reflected laser light would be collected by the receiver. The key parameters specifying system performance are

Resolution, d
Number of resolution elements in image, N by N
RMS error in the image, e .

The use of the reduced tolerance imaging concept offers (1) the possibility of attaining desired values of these parameters at lower cost, lower weight, less complexity of the optical system, and less disruption of aircraft aerodynamics and (2) the possibility of diffraction-limited imaging through atmospheric or aircraft turbulence.

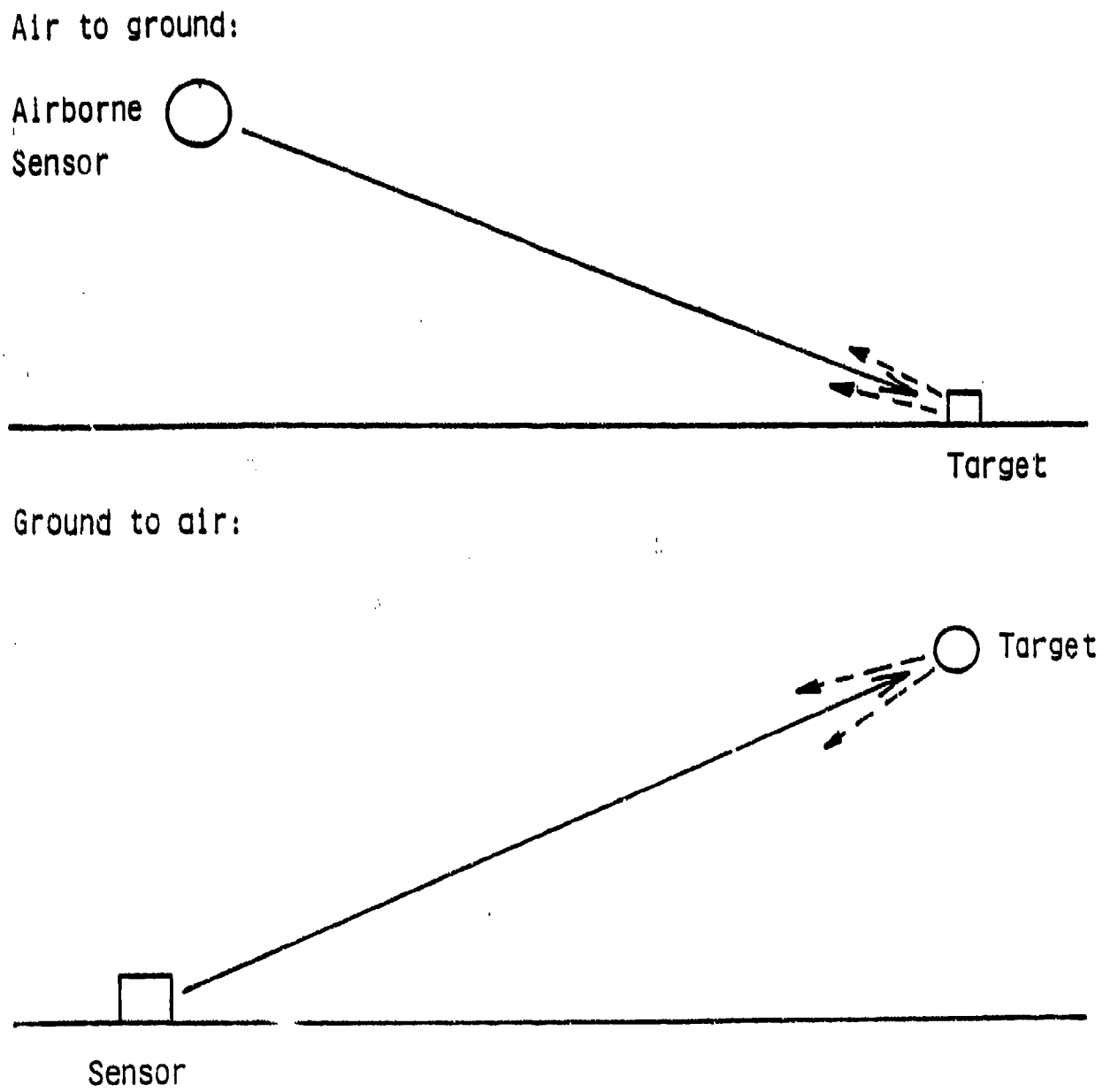


FIGURE 9-1. TACTICAL IMAGING APPLICATION USING LASER ILLUMINATION.

9.3 SYSTEM ARCHITECTURE

To use the reduced tolerance imaging concept, the imaging system will have optical and data processing architectures that are significantly different from those of conventional imaging systems. The transmitting and receiving optics are shown in Fig. 9-2. A pulsed laser beam is spatially filtered, expanded, and passed through a mask. The mask is imaged to the target region to provide an illumination-pattern-determined support constraint. For efficient illumination of the mask (which might consist of a transmitting region of triangular shape), holographic beamforming elements should be considered. To produce an image of the mask which has low sidelobes, apodization of the imaging system is necessary. Low sidelobes are required to minimize the difficulties encountered in reconstructing images using tapered support constraints. Although a lens is used to image the mask in Fig. 9-2, reflective optics could also be used.

Two versions of the receiver optics are shown in Fig. 9-2. Both make far-field intensity measurements of the laser light reflected from the target region. The conformal detector consists of an appropriate photon-to-electron conversion material applied conformally to the surface of a vehicle (for example, to the fuselage of an aircraft). If this type of detector were to be developed, it would certainly be the lightest, least complex, and least disruptive of aerodynamics of any detector approach. The other detector approach shown in Fig. 9-2 uses a lenslet array to focus the light in each subaperture of the receiver onto a single detector. Although heavier and more complex than the conformal detector, this detector design is achievable with current technology, permits greater control of the field of view of the receiver, and should not significantly affect aerodynamics. Either detector will use a narrow-band filter to eliminate light at wavelengths other than that of the laser and a polarizer to select only a single polarization. It should be noted that neither detector is affected by turbulence which appears as a phase aberration in the receiver aperture.

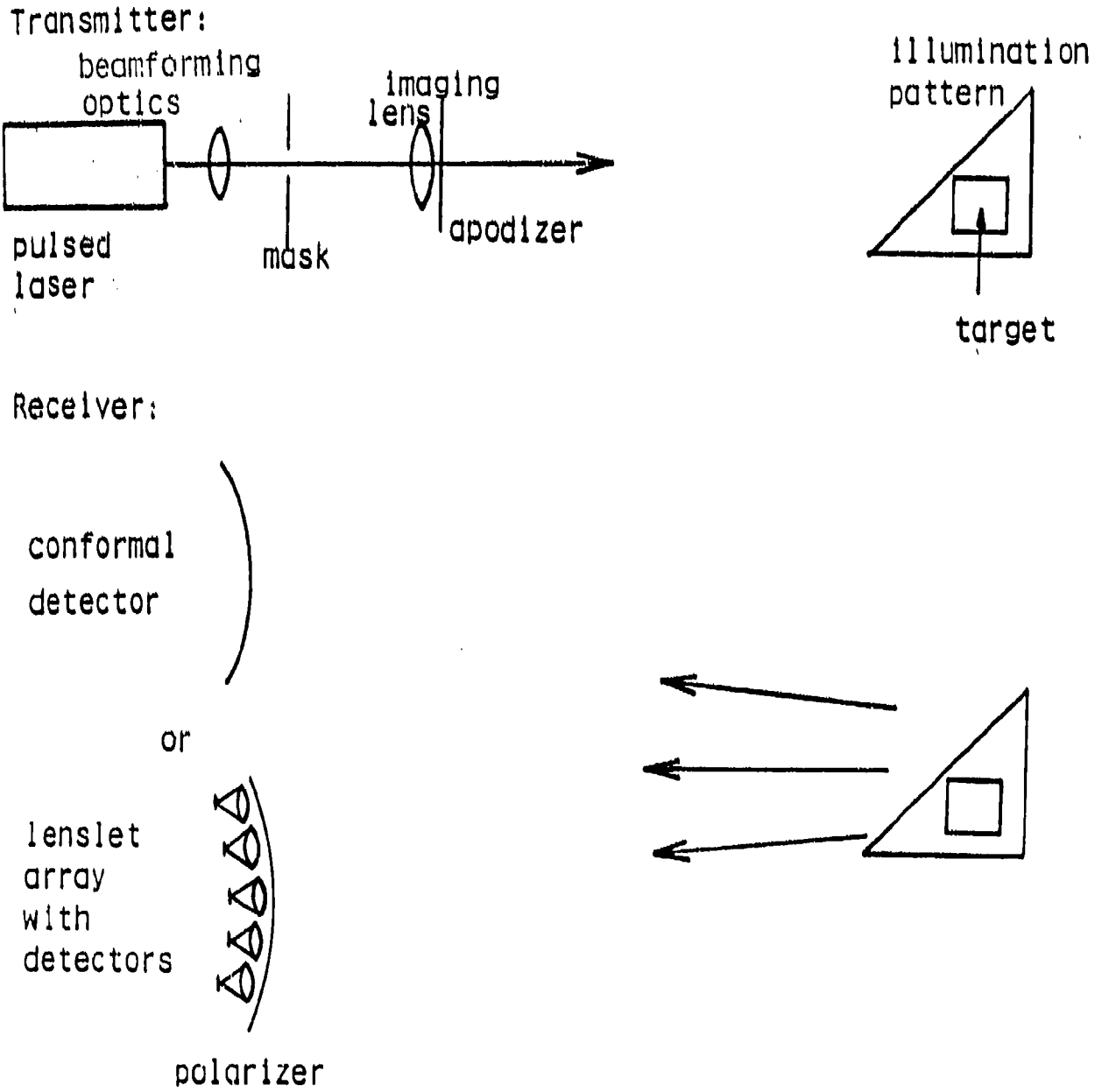


FIGURE 9-2. TRANSMITTING AND RECEIVING OPTICS FOR REDUCED TOLERANCE IMAGING.

A block diagram of the data processing architecture is shown in Fig. 9-3. The far-field intensity data from the receiver is first preprocessed to reduce detector noise, background bias, fixed pattern effects, and nonlinear and nonuniform detector response effects. The resulting data is then input to the iterative Fourier transform algorithm along with the image support constraint as determined by the illumination pattern. Within the iterative Fourier transform algorithm, the error reduction, hybrid input-output, and enlarging mask options would be used. The image reconstructed by the algorithm could be postprocessed to reduce remaining noise or to reduce speckle, if desired.

The hardware required to perform this data processing will vary with the time allowed to produce the reconstructed image. The dominant computational operation is the fast Fourier transform (FFT) required by the iterative algorithm. To reconstruct an N by N image, a 2N by 2N FFT is required which in turn requires $8N^2\log_2 N$ floating point multiplies. There are 2 FFTs per iteration, so, for a reconstruction requiring M iterations, the number of multiplies required is $16MN^2\log_2 N$. The other computations required (such as constraint enforcement, pre- and post-processing) are small in comparison. For M=100 and N=128, the total number of multiplies required for a reconstruction is about 184×10^6. Current or near-term single VHSIC chips should achieve 40 megaflop computation rates, so a small special-purpose processor utilizing as few as 5 VHSIC chips could reconstruct a 128 by 128 image in less than a second.

9.4 SYSTEM DESIGN PARAMETERS

This section lists the parameters which must be determined in a system design and derives important relationships between these parameters and the system performance parameters defined in Sec. 9.2. The key parameters in system design are

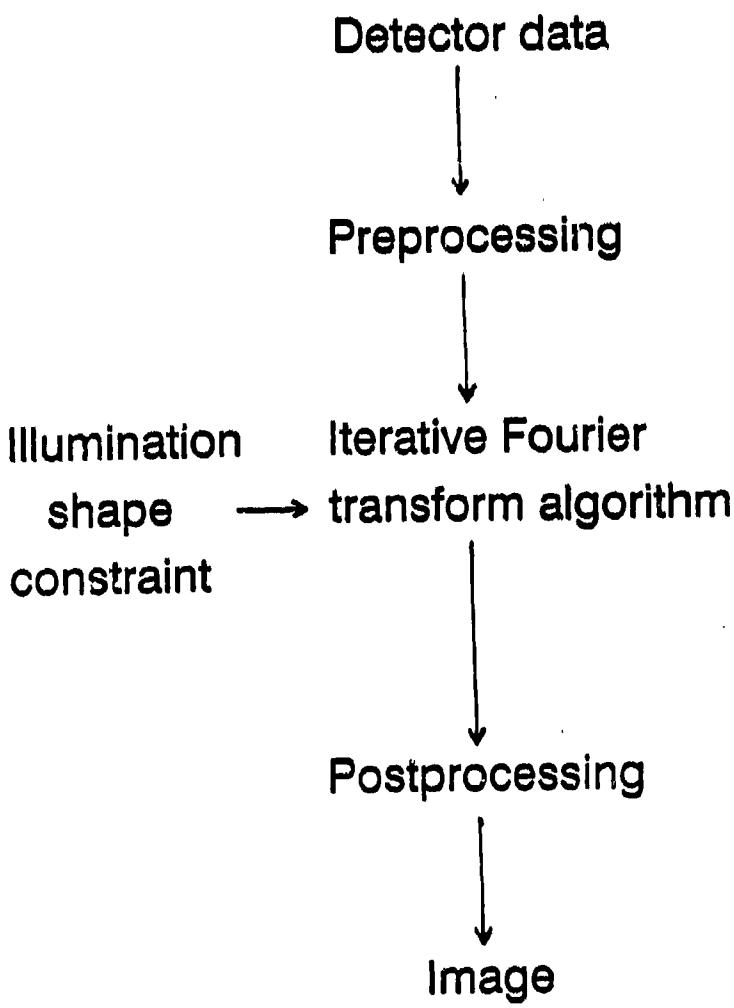


FIGURE 9-3. DATA PROCESSING ARCHITECTURE.

Transmitter:

- Laser wavelength, λ
- Laser pulse energy, E_t
- Laser pulse length, $\Delta\tau$
- Laser spatial coherence
- Laser temporal coherence
- Transmitter aperture diameter, D_t
- Transmitter optical efficiency, t_t

Optical path:

- Two-way path transmittance, $\exp(-2\alpha R)$

Target:

- Target range, R
- Target reflectivity, ρ
- Illumination pattern taper
  in image resolution elements, n

Receiver:

- Receiver aperture diameter, D_r
- Receiver optical transmittance, t_r
- Number of detector elements, 2N by 2N
- Detector quantum efficiency, η
- Detector thermal and readout noise
  standard deviation in electrons, n_t

9.4.1 ILLUMINATION PATTERN TAPER

One important relationship between some of these parameters concerns illumination pattern taper. The size of a resolution element in the illumination pattern at the target is approximately $1.5\lambda R/D_t$. The factor of 1.5 is present because the aperture is apodized to reduce

sidelobes. Since the resolution, d, of the image (measured at the object) is given by

$$d = \lambda R/D_r,$$

the taper, n, in image resolution elements is

$$n = 1.5D_r/D_t .$$

It is desirable for the taper to be small since then the image reconstruction will have lower error for a given measurement signal-to-noise ratio (SNR). It is also desirable for the transmitter aperture to be small since then the transmitter optics will be lighter, less disruptive of aerodynamics, and less affected by turbulence. For a fixed receiver aperture, the above relationship indicates that a design tradeoff must be made between taper and transmitter aperture.

9.4.2 DETECTOR ELEMENT COLLECTION AREA

Because the far-field intensity to be measured has twice the bandwidth of the far-field complex amplitude, the number of measurements required is 2N by 2N to enable reconstruction of an N by N image. The far-field intensity is a speckle pattern with each speckle approximately of diameter $\lambda R/Nd$ where Nd is the diameter of the illuminated target region. To achieve the necessary Nyquist sampling of the speckle pattern, the sample spacing must be equal to half the speckle diameter. Each detector element, whether of the conformal or lenslet array type, must therefore collect the light over a region no larger in diameter than $\lambda R/2Nd$.

If it is assumed that the target region scatters light uniformly over 2π steradians, then the fraction of the reflected light collected by each detector element (collection efficiency) is given by the ratio

of the collection area (of a detector element) to $2\pi R^2$. Using the diameter $\lambda R/2Nd$ given in the previous paragraph, the result, assuming a circular collection area, is

$$\text{collection efficiency} = (\lambda/Nd)^2/32 \ .$$

It is important to note that although collection efficiency does depend on the size of the region to be imaged, it does not explicitly depend on target range as long as the detector elements are sized according to the range to the target.

9.4.3 RECEIVED ENERGY

The optical energy, E_r, received by a single detector element is the product of the laser pulse energy, the transmitter optical efficiency, the two-way path transmittance, the target reflectivity, the detector collection efficiency, and the receiver optical efficiency. Using the notation defined at the beginning of Sec. 9.4, the received optical energy is

$$E_r = E_t \ t_t \ t_r \ r(\lambda/Nd)^2 \exp(-2\alpha R)/32 \ .$$

Given that the parameters t_t, t_r, r, α, and R are either determined by the application or will be made as large as possible, this equation states that, to achieve a desired E_r (and therefore a desired SNR and RMS image error), a tradeoff must be made between laser pulse energy, E_t, and the area, $(Nd)^2$, of the region to be imaged.

9.4.4 MEASUREMENT SIGNAL-TO-NOISE RATIO

Assuming that the measured data can be corrected for any background bias, fixed pattern effects, and nonlinear and nonuniform detector response effects, then the primary remaining degredations of the data will be due to the Poisson statistics of the received photons and the thermal noise of the detector. The mean number of signal photoelectrons, n_s, generated per detector element by the received energy is

$$n_s = \eta E_r/(hc/\lambda)$$

where η is the detector quantum efficiency; the Planck constant, h, equals 6.63×10^{-34} Joule sec; and the speed of light, c, equals 3×10^8 m/sec. The standard deviation, n_p, of the number of photoelectrons, for Poisson statistics, is

$$n_p = \sqrt{n_s} .$$

The SNR of the detected far-field intensity data is

$$SNR = \frac{n_s}{\sqrt{n_p^2 + n_t^2}}$$

where n_t is the standard deviation of the detector thermal and readout noise. Depending on the wavelength region and the detector used, either Poisson or thermal noise may dominate the denominator of this expression.

9.4.5 LASER PULSE LENGTH AND COHERENCE

The ability to reconstruct an image from far-field intensity measurements depends on the far-field complex amplitude accurately representing the Fourier transform of the complex-valued reflectivity of

the illuminated region. This in turn implies that the laser spatial coherence must equal or exceed the width of the illuminated region and that the temporal coherence must equal or exceed the depth (along the line-of-sight from the imaging system) of the region. For the case of a target located on the ground, the depth of the illuminated region will depend on the slant angle of the illumination. The temporal coherence of the laser will therefore determine the maximum illumination angle relative to the normal to the ground plane.

In addition to sufficient coherence, for a pulsed laser, the pulse length must be much greater than the depth of the illuminated region. This requirement ensures that, for most of the pulse length, light will be arriving at the receiver from all parts of the target region. If this is not the case, then, as mentioned in the previous paragraph, the far-field complex amplitude will not represent a Fourier transform of the target.

9.5 NUMERICAL EXAMPLE OF A SYSTEM DESIGN

For a numerical example of a system design, the system performance parameters were chosen to be

Resolution, $d = 0.1$ m
Number of resolution elements
in image, N by N = 128 by 128
RMS error in the image, $e = 10\%$.

The following system design parameters were chosen or assumed:

Transmitter:

Laser wavelength, $\lambda = 10.6\ \mu m$
Laser pulse length, $\Delta\tau = 1\ \mu sec$

Laser spatial and temporal coherence $>$ 100 m
(these requirements, which are easily achieved by a CO_2 laser, assure that a cube 12.8 meters on a side can be imaged)
Transmitter optical efficiency, $t_t = 0.5$

Target:

Target range, R = 10 km
Target reflectivity, r = 0.1
Illumination pattern taper
in image resolution elements, n = 6

Optical path:

Two-way path transmittance, exp(-2) = 0.135
(α = 0.1/km)

Receiver:

Receiver optical efficiency, $t_r = 0.5$
Number of detector elements, 256 by 256
Detector type: lenslet array
Detector quantum efficiency, $\eta = 0.5$
Detector $D^* = 1 \times 10^{11}$ cm$(Hz)^{1/2}$/Watt
(HgCdTe at 77°K with an f/4 lenslet array)

From the equations derived earlier, the following dependent parameters are then determined:

Transmitter aperture diameter, $D_t = 0.26$ m
Receiver aperture diameter, $D_r = 1.06$ m
Detector collection area diameter, $D_r/2N = 4$ mm .

For an image error of 10% and an illumination pattern taper of 6 image resolution elements, the required SNR in the far-field intensity data

has been shown by computer simulations to about 10 to 1. For infrared detectors, the dominant noise is thermal. Using the definition of D^*, the noise equivalent received energy, E_n, is

$$E_n = \frac{\Delta w \sqrt{(\Delta \tau / 2)}}{D^*}$$

Assuming the detector width, Δw, is 50 μm, then $E_n = 3.5 \times 10^{-17}$ Joules. This corresponds to a thermal noise standard deviation, n_t, of 940 electrons. The laser pulse energy, E_t, required to produce a mean signal 10 times as large (for a SNR of 10 to 1) is 5 Joules. The standard deviation, n_p, of the mean signal due to Poisson statistics is 100 electrons which, as expected, is much smaller than n_t.

9.6 CONCLUSION

The practicality of the reduced tolerance imaging concept for tactical air-to-ground or ground-to-air imaging using active laser illumination has been investigated. The key system performance parameters were defined and transmitter, receiver, and data processing algorithm architectures were established. Relationships were determined between these performance parameters and system, target, and optical path parameters. The main design tradeoffs identified were transmitter aperture diameter versus illumination pattern taper, and laser pulse energy versus the size of the region to be imaged. The practicality of an imaging system at 10.6 μm was explored through computation of numerical values for the system parameters using realistic values for system performance, optical path and target characteristics, and infrared detector performance. For the example given, the largest coherent aperture required (for the transmitter) was only one-fourth the diameter (one-sixteenth the area) of the coherent aperture we would have needed to form a diffraction-limited image. Based on this analysis, reduced tolerance imaging continues to offer the possibility of reducing cost, weight, complexity, and aerodynamic effect of the imaging system and also the possiblility of diffraction-limited imaging through turbulence.

Appendix A

PARAMETER ESTIMATION AND THE CRAMER-RAO LOWER BOUND

A.1 INTRODUCTION

A.1.1 The Estimation Problem

The problem we consider is that of estimating a parameter vector $\underline{a} = (a_1, \ldots, a_L)$ from a measurement vector $\underline{R} = (R_1, \ldots, R_M)$. An *estimate* is a random vector $\hat{\underline{A}} = (\hat{A}_1, \ldots, \hat{A}_L)$, where each $\hat{A}_i$ is intended as an approximation to the corresponding a_i. An *estimator* is a function $\hat{\underline{A}}(\cdot)$ mapping measurements into parameter values. The *estimator* $\hat{\underline{A}}(\cdot)$ *produces* the *estimate* $\hat{\underline{A}} = \hat{\underline{A}}(\underline{R})$ of $\underline{a}$. It is assumed that the conditional probability density $p(\underline{r} \mid \underline{a})$ is known.

A word on notation is in order. In general, we will use upper case letters to denote random variables and vectors and will use lower case letters to denote their possible values and also most non-random variables and vectors. Vectors will be underscored. Thus, the received vector, which contains randomness, is denoted $\underline{R}$. A function of a random quantity is also to be considered random; hence, the estimate $\hat{\underline{A}}$ is capitalized. As described below, the parameter vector to be estimated may be random or not. In the previous paragraph, we elected to use the lower case $\underline{a}$. The probability density of a random vector such as $\underline{R}$ will be denoted $p_{\underline{R}}(\underline{r})$, or simply $p(\underline{r})$, when no confusion can arise. A conditional

probability density such as that of $\underline{R}$ given $\underline{A}$ will be denoted $p_{\underline{R}|\underline{A}}(\underline{r}\,|\,\underline{a})$ or $p(\underline{r}\,|\,\underline{a})$. The average of a random quantity will be denoted either $E[A]$ or $E_A[a]$. The latter is especially helpful for expressions like $E_A[dp_A(a)/da]$, which otherwise would have to be written $E[dp_A(A)/dA]$.

We consider two distinct estimation environments that differ primarily in their prior knowledge of the parameter vector $\underline{a}$ to be estimated.

Environment 1: Nonrandom Parameters

Here, the parameter vector to be estimated is a fixed but unknown vector $\underline{a}$, and the quality of an estimator $\hat{A}(\cdot)$ is judged by the resulting *mean squared error functions*

$$S_i(\underline{a}) \triangleq E[(\hat{A}_i - a_i)^2], \quad i = 1, 2, \ldots, L \quad . \tag{A-1.1}$$

As these mean squared errors ordinarily depend on the actual values of the parameters $\underline{a}$, there is not usually a "best" estimator. Rather some estimators are better for certain values of $\underline{a}$ and worse for others.

A frequently used estimator for this environment is the *maximum likelihood* (ML) estimator, which chooses $\hat{A}(\underline{r}) = \underline{a}$ if $p(\underline{r}\,|\,\underline{a})$ is largest among all choices for $\underline{a}$.

An estimator is *unbiased* if $E[\hat{A}_i] = a_i$, $i = 1, \ldots, L$. The ML estimator might or might not be unbiased. In some special cases there exists an unbiased

estimator $\hat{A}^*(\cdot)$ that is *best* in the sense that for any other estimator $\hat{A}$,

$$S_i^*(a) \triangleq E[(\hat{A}^*(R) - a_i)^2] \leq S_i(a) \triangleq E[(\hat{A}(R) - a_i)^2], \quad \text{for all } a \text{ and } i = 1, \ldots, L. \tag{A-1.2}$$

Environment 2: Random Parameters

Here, the parameter vector to be estimated is a random vector A with known probability density $p(a)$. The quality of an estimator is judged by the *mean squared errors*

$$S_i \triangleq E[(\hat{A}_i - A_i)^2], \quad i = 1, \ldots, L. \tag{A-1.3}$$

In this environment, there is a best estimator, namely, the *minimum mean squared error* (MMSE) estimator:

$$\hat{A}(r) = E[A \mid R = r]. \tag{A-1.4}$$

An estimator in Environment 2 is said to be *unbiased* if $E[\hat{A}_i] = E[A_i]$, $i = 1, \ldots, L$. The MMSE estimator is unbiased.

A.1.2. The Cramer-Rao Lower Bound

In either environment, there is a Cramer-Rao lower bound (CRLB) to the MSE.

Environment 1: For any unbiased estimator,

$$E[(\hat{A}_i - a_i)^2] \geq [J]^{-1}_{ii} \quad , \quad i = 1, \ldots, L \tag{A-1.5}$$

where J is the $L \times L$ matrix, sometimes called the Fisher information matrix, with

$$J_{ij} = E_{\underline{R}} \left[\frac{\partial}{\partial a_i} \ln p(\underline{r} \mid \underline{a}) \frac{\partial}{\partial a_j} \ln p(\underline{r} \mid \underline{a}) \right] \tag{A-1.6a}$$

$$= - E_{\underline{R}} \left[\frac{\partial^2}{\partial a_i \partial a_j} \ln p(\underline{r} \mid \underline{a}) \right] \quad , \tag{A-1.6b}$$

and where the expectations average over $\underline{R}$. The Cramer-Rao bound also gives the more complete result

$$C \geq J^{-1} \quad , \tag{A-1.7}$$

where $C = [C_{ij}]$ is the covariance matrix of the errors ($C_{ij} = E[(\hat{A}_i - a_i)(\hat{A}_j - a_j)]$ and where $C \geq J^{-1}$ means that $C - J^{-1}$ is a non-negative definite matrix. (A matrix M is non-negative definite if $\underline{x}^t M \underline{x} \geq 0$, for every column vector $\underline{x}$). Note that $C \geq J^{-1}$ implies $C_{ii} \geq [J]^{-1}_{ii}$, so (A-1.7) subsumes (A-1.5).

Environment 2: For estimators having a *certain "unbiased-like"* property,

$$E[(\hat{A}_i - A_i)^2] \geq [\bar{J} + K]^{-1}_{ii} \quad , \quad i = 1, \ldots, L \quad , \tag{A-1.8}$$

where

$$\bar{J}_{ij} = E_{\underline{A}}[J_{ij}] \tag{A-1.9}$$

and

$$K_{ij} = E_{\underline{A}} \left[\frac{\partial}{\partial a_i} \ln p(\underline{a}) \frac{\partial}{\partial a_j} \ln p(\underline{a}) \right] \qquad \text{(A-1.10a)}$$

$$= - E_{\underline{A}} \left[\frac{\partial^2}{\partial a_i \partial a_j} \ln p(\underline{a}) \right] . \qquad \text{(A-1.10b)}$$

In this environment, the expectations are average over $\underline{R}$ and $\underline{A}$. In addition there is also the more complete result:

$$C \geq [\bar{J} + K]^{-1} \qquad \text{(A-1.11)}$$

where $C = [C_{ij}]$ is the covariance matrix of the errors, $(C_{ij} = E[(\hat{A}_i - A_i)(\hat{A}_j - A_j)])$. In the scalar case (i.e., $L = 1$) the "unbiased-like" property which the estimator must satisfy is

$$\lim_{a \to \pm\infty} p_A(a) \left(E_{\underline{R}} [\hat{A}(\underline{r}) \mid A_i = a] - a \right) = 0 \quad . \qquad \text{(A-1.12)}$$

Notes:

(1) The existence of the above derivatives is presumed.

(2) The matrices J and K are functions of the likelihood distribution $p(\underline{r} \mid \underline{a})$ and the a priori distribution $p(\underline{a})$, respectively. $\bar{J}$ is related primarily to the likelihood distribution $p(\underline{r} \mid \underline{a})$ and secondarily to the a priori distribution $p(\underline{a})$.

A.2. SCALAR PARAMETERS

A.2.1. In this section we consider the case where the parameter to be estimated is a scalar A (i.e., $L = 1$). The measurements may be a vector or a scalar. In this case the Cramer-Rao lower bounds simplify as shown below.

Environment 1:

$$E[(\hat{A} - a)^2] \geq \frac{1}{J} \tag{A-2.1}$$

where

$$J = -E_R\left[\frac{\partial^2}{\partial a^2} \ln p(\underline{r} \mid a)\right] = E_R\left[\left(\frac{\partial}{\partial a} \ln p(\underline{r} \mid a)\right)^2\right] . \tag{A-2.2}$$

Environment 2:

$$E[(\hat{A} - A)^2] \geq \frac{1}{J+K} \tag{A-2.3}$$

where

$$\bar{J} = E_A[J] = -E_{R,A}\left[\frac{\partial^2}{\partial a^2} \ln p(\underline{r} \mid a)\right] = E_{R,A}\left[\left(\frac{\partial}{\partial a} \ln p(\underline{r} \mid a)\right)^2\right] \tag{A-2.4}$$

$$K = -E_A\left[\frac{\partial^2}{\partial a^2} \ln p(a)\right] = E_A\left[\left(\frac{\partial}{\partial a} \ln p(a)\right)^2\right] = E_A\left[\left(\frac{p'_A(a)}{p_A(a)}\right)^2\right] , \tag{A-2.5}$$

where $p'_A(a)$ denotes the derivative of $p_A(a)$ with respect to a.

Note that if in Environment 2 one has to estimate a parameter vector $A = (A_1, ..., A_L)$ from a measurement vector R, one may separately apply the bound in (A-2.3) to each component of A and obtain

$$E[(\hat{A}_j - A_j)^2] \geq \left(E_{R,A_j}\left[\left(\frac{\partial}{\partial a_j} \ln p(R \mid a_j) \right)^2 \right] + E_{A_j}\left[\left(\frac{\partial}{\partial a_j} p(a_j) \right)^2 \right] \right)^{-1} \tag{A-2.6}$$

This bound is presumably simpler to compute but not as good as that of Eq. (A-1.7).

A.2.2 Additive noise problems: Calculation of J

(a) Suppose

$$R = f(A) + N \quad , \tag{A-2.7}$$

where f is some function and N is some noise that is independent of A with density $p_N(n)$. Then for estimating A from R,

$$J = E_N\left[\left(\frac{p'_N(n)}{p_N(n)} \right)^2 \right] (f'(a))^2 \quad , \tag{A-2.8}$$

where $p'_N(n)$ and $f'(a)$ denote the derivatives of $p_N(n)$ and $f(a)$ with respect to n and a, respectively.

(b) Suppose (A-2.7) holds and N is Gaussian with mean m_N and variance σ_N^2, i.e., $\eta(m_N, \sigma_N^2)$. Then

$$J = \frac{1}{\sigma_N^2}\left[f'(a)\right]^2 \tag{A-2.9}$$

(c) Suppose (A-2.7) holds with N Gaussian $\eta(m_N, \sigma_N^2)$ and with

$$f(a) = ca + d \quad . \tag{A-2.10}$$

Then

$$J = \frac{c^2}{\sigma_N^2} \tag{A-2.11}$$

Notice that J does not depend on a. For this case the ML estimator is

$$\hat{A}(r) = \frac{r-d}{c} \quad , \tag{A-2.12}$$

which is unbiased and results in mean squared error

$$E\left[(\hat{A}(R)-a)^2\right] = \frac{c^2}{\sigma_N^2} \quad , \tag{A-2.13}$$

which in turn coincides with 1/J, the Cramer-Rao lower bound to mean squared error. Thus, we draw two conclusions: the ML estimator is optimum in the sense that no unbiased estimator has smaller mean squared error, and the Cramer-Rao bound is tight.

A.2.3 Environment 2: Calculation of K

If A is Gaussian $\eta(m_A, \sigma_A^2)$, then

$$K = \frac{1}{\sigma_A^2} \quad . \tag{A-2.14}$$

A.2.4 Linear Gaussian Problem: Environment 2

Suppose

$$R = cA + d + N \quad , \tag{A-2.15}$$

where A and N are independent, Gaussian, $\eta(m_A, \sigma_A^2)$ and $\eta(m_N, \sigma_N^2)$, respectively. In this situation, the Cramer-Rao lower bound again turns out to be tight, i.e., it equals the mean squared error of the best estimator. Specifically, the best estimator $\hat{A}^*(R) = E[A \mid R]$ turns out to be the ML estimator

$$\hat{A}^*(R) = \frac{R-d}{c} \quad , \tag{A-2.16}$$

and the resulting mean squared error is

$$E[(\hat{A}^* - A)^2] = \frac{1}{J+K} = \frac{1}{\frac{c^2}{\sigma_N^2} + \frac{1}{\sigma_A^2}} = \frac{\sigma_A^2}{\frac{c^2\sigma_A^2}{\sigma_N^2} + 1} \quad . \tag{A-2.17}$$

A.3 INDEPENDENT MEASUREMENTS: THE J-MATRIX

A.3.1 Conditionally Independent Measurements

Let us consider the commonly occurring situation wherein the measurements $R_1, \ldots, R_M$ are conditionally independent given the parameter vector $\underline{A}$. That is,

$$p_{\underline{R}|\underline{A}}(\underline{r}\,|\,\underline{a}) = \prod_{k=1}^{M} p_k(r_k\,|\,\underline{a}) \quad , \tag{A-3.1}$$

where for brevity we write $p_k(r_k\,|\,\underline{a})$ for $p_{R_k|\underline{A}}(r_k\,|\,\underline{a})$. We may think of each R_k as an independent measurement of $\underline{A}$. For example, such conditional independence holds if

$$R_k = f_k(\underline{A}, N_k) \quad , \quad k = 1, \ldots, M \quad , \tag{A-3.2}$$

where $f_1, \ldots, f_M$ are known functions and $N_1, \ldots, N_M$ are random variables that are independent of each other and of $\underline{A}$.

Then as shown in Section A.3.3, the J-matrix in the Cramer-Rao bound for estimating $\underline{A}$ based on $\underline{R}$ is

$$J = \sum_{k=1}^{M} J^{(k)} \quad , \tag{A-3.3}$$

where $J^{(k)}$ is the J-matrix for the Cramer-Rao lower bound to the MSE for estimating $\underline{A}$ based on R_k; that is,

$$J_{ij}^{(k)} = E_{R_k}\left[\frac{\partial}{\partial a_i}\ln p_k(r_k \mid \underline{a})\,\frac{\partial}{\partial a_j}\ln p_k(r_k \mid \underline{a})\right] . \qquad \text{(A-3.4)}$$

Equation (A-3.3) shows that each independent measurement R_k has an additive effect on J. The K-matrix (in Environment 2) has the usual form.

One specific example occurs frequently, namely,

$$R_k = f_k(A) + N_k \ , \quad k = 1, \ldots, M \ , \qquad \text{(A-3.5)}$$

where $f_1, \ldots, f_M$ are known functions and $N_1, \ldots, N_M$ are independent of A and each other. Then using the fact that

$$p_k(r_k \mid \underline{a}) = p_{N_k}(r_k - f_k(\underline{a})) \ , \qquad \text{(A-3.6)}$$

one finds

$$J_{ij}^{(k)} = E_{N_k}\left[\left(\frac{p'_{N_k}(n_k)}{p_{N_k}(n_k)}\right)^2\right]\frac{\partial}{\partial a_i}\,f_k(\underline{a})\,\frac{\partial}{\partial a_j}\,f_k(\underline{a}) \ . \qquad \text{(A-3.7)}$$

If in addition N_k is Gaussian $\eta(m_{N_k}, \sigma_{N_k}^2)$, then

$$J_{ij}^{(k)} = \frac{1}{\sigma_{N_k}^2}\,\frac{\partial}{\partial a_i}\,f_k(\underline{a})\,\frac{\partial}{\partial a_j}\,f_k(\underline{a}) \ . \qquad \text{(A-3.8)}$$

A.3.2 One-for-one Independent Measurements

Let us here consider the situation, henceforth referred to as *one-for-one independent measurements*, where there are equal numbers of measurements and parameters to estimate (i.e., $L = M$) and where in addition to (A-3.1) we have

$$p_{R\,|\,A}(r\,|\,a) = \prod_{k=1}^{M} p_k(r_k\,|\,a_k) \quad , \qquad \text{(A-3.9a)}$$

or equivalently,

$$p_{R_k\,|\,A}(r_k\,|\,a) = p_k(r_k\,|\,a_k) \, , \qquad \text{(A-3.9b)}$$

where $p_k(r_k\,|\,a_k) \triangleq p_{R_k\,|\,A_k}(r_k\,|\,a_k) = p_{R_k\,|\,A}(r_k\,|\,a)$. That is, each measurement R_k depends only on the corresponding parameter A_k. Therefore, we may think of each R_k as an independent measurement of A_k. This situation occurs, for example, when

$$R_k = f_k(A_k, N_k) \quad , \quad k = 1, \ldots, M \quad , \qquad \text{(A-3.10)}$$

for some functions $f_1, \ldots, f_M$ and random variables $N_1, \ldots, N_M$ that are independent of each other and of A.

Environment 1.

In the case of one-for-one independent measurements in Environment 1, each parameter A_k may without loss of optimality be estimated from R_k alone; i.e., without using $R_j, j \neq k$, and without regard to the joint distribution of $(A_1, \ldots, A_M)$. We will now show that the Cramer-Rao lower bound reflects this fact by reducing to the scalar lower bound.

It follows from (A-3.3), (A-3.4) and (A-3.9b) that $J_{ij}^{(k)} = 0$ unless $i = j = k$. Hence, J is diagonal. Specifically,

$$J = \begin{bmatrix} J_{11} & & & \\ & J_{22} & 0 & \\ & & \ddots & \\ 0 & & & J_{MM} \end{bmatrix} \qquad \text{(A-3.11)}$$

where

$$J_{kk} = J_k^{(k)} = E_{R_k}\left[\left(\frac{\partial}{\partial a_k} \ln p_k(r_k \mid a_k)\right)^2\right] . \qquad \text{(A-3.12)}$$

Substituting (A-3.11) into (A-1.5) shows

$$E[\hat{A}_k - a_k]^2 \geq \frac{1}{J_{kk}} , \qquad \text{(A-3.13)}$$

which is exactly the same bound that one would get by applying the scalar lower bound (A-2.1) to the problem of estimating A_k from R_k alone.

One commonly occurring instance of one-for-one independent measurements is where

$$R_k = f_k(A_k) + N_k , \quad k = 1, \ldots, M , \qquad \text{(A-3.14)}$$

where the F_k's are known functions and the N_k's are independent of each other and of A. In this case

$$p_k(r_k \mid a_k) = p_{N_k}(r_k - f_k(a_k)) . \qquad \text{(A-3.15)}$$

Substituting the above into (A-3.12), gives (alternatively, it follows from (A-2.8))

$$J_{kk} = E_{N_k}\left[\left(\frac{p'_{N_k}(n_k)}{p_{N_k}(n_k)}\right)^2\right]\left(f'_k(a_k)\right)^2 \qquad \text{(A-3.16)}$$

Finally, if the N_k's are Gaussian $\eta(m_{N_k},\sigma^2_{N_k})$, then

$$J_{kk} = \frac{1}{\sigma^2_{N_k}}\left(f'_k(a_k)\right)^2 \quad . \qquad \text{(A-3.17)}$$

Environment 2

Consider the situation of one-for-one independent measurements in Environment 2. If in addition the components of $\underline{A} = [A_1,...,A_M]$ are independent, then without loss of optimality one may estimate each A_k from R_k alone. That is the optimum estimate of A_k is

$$\hat{A}_k = E[A_k \mid \underline{R}] = E[A_k \mid R_k] \quad . \qquad \text{(A-3.18)}$$

We now show that the Cramer-Rao lower bound reflects this fact by reducing to the scalar lower bound.

It follows from (A-3.11) that $\bar{J} = E_{\underline{A}}[J]$ is a diagonal matrix with

$$\bar{J}_{kk} = E_{R_k A_k}\left[\left(\frac{\partial}{\partial a_k}\ln p(r_k \mid a_k)\right)^2\right] \qquad \text{(A-3.19)}$$

In addition, the independence of $A_1,...,A_M$ imply that the K matrix is diagonal with

$$K_{kk} = E_{A_k}\left[\left(\frac{\partial}{\partial a_k}\ln p(a_k)\right)^2\right] \quad . \qquad \text{(A-3.20)}$$

Thus $[\bar{J} + K]$ and $[\bar{J} + K]^{-1}$ are diagonal matrices and one obtains the Cramer-

Rao lower bound

$$E\left[\left(\hat{A}_k - A_k\right)^2\right] \geq \left[\bar{J} + K\right]^{-1}_{kk} \tag{A-3.21}$$
$$= E_{R_k A_k}\left[\left(\frac{\partial}{\partial a_k} \ln p(r_k \mid a_k)\right)^2\right] + E_{A_k}\left[\left(\frac{\partial}{\partial a_k} \ln p(a_k)\right)^2\right],$$

which is exactly the same bound that one would get by applying the scalar lower bound (A-2.3) to the problem of estimating A_k from R_k alone.

If, however, the components of $\underline{A}$ are *dependent*, then the best estimate of each A_k will depend on other R_j's besides R_k. Intuitively, we may think that R_j, $j \neq k$, contains information about A_j, which in turn has information about A_k. This dependence is reflected in the Cramer-Rao lower bound. Specifically, the matrix K and, consequently, $[\bar{J} + K]$ will not be diagonal. Note, however, that $\bar{J}$ will be diagonal, and this sometimes allows some simplification.

A.3.3 Proof of (A-3.3)

Recall that

$$J_{ij} \triangleq E\left[\frac{\partial}{\partial a_i} \ln p(\underline{R} \mid \underline{a}) \frac{\partial}{\partial a_j} \ln p(\underline{R} \mid \underline{a})\right] . \tag{A-3.22}$$

Using (A-3.1), we find

$$\frac{\partial}{\partial a_i} \ln p(\underline{R} \mid \underline{a}) = \frac{\partial}{\partial a_i} \sum_{k=1}^{N} \ln p_k(R_k \mid \underline{a}) = \sum_{k=1}^{N} Y_{i,k} , \tag{A-3.23}$$

where $Y_{i,k}$ is the random variable

$$Y_{i,k} \triangleq \frac{\partial}{\partial a_i} \ln p_k(R_k \mid \underline{a}) \quad . \tag{A-3.24}$$

The assumption of conditionally independent measurements implies that $Y_{i,k}$ is independent of $Y_{j,l}$, if $k \neq l$. Substituting (A-3.23) into (A-3.22) and using the independence of the $Y_{i,k}$'s gives

$$\begin{aligned} J_{ij} &= E\left[\sum_{k=1}^{N} Y_{i,k} \sum_{l=1}^{N} Y_{j,l}\right] = \sum_{k=1}^{N} \sum_{l=1}^{N} E[Y_{i,k}\ Y_{j,l}] \\ &= \sum_{k=1}^{N} E[Y_{i,k}\ Y_{j,k}] + \sum_{k=1}^{N} \sum_{\substack{l=1 \\ l \neq j}}^{N} E[Y_{i,k}]\, E[Y_{j,l}] \quad . \end{aligned} \tag{A-3.25}$$

We now show that the second term in the above is zero. For any i,k

$$\begin{aligned} E[Y_{i,k}] &= \int \left[\frac{\partial}{\partial a_i} \ln p_k(r \mid \underline{a})\right] p_k(r \mid \underline{a})\, dr = \int \left[\frac{\frac{\partial}{\partial a_i} p_k(r \mid \underline{a})}{p_k(r \mid \underline{a})}\right] p_k(r \mid \underline{a})\, dr \\ &= \int \frac{\partial}{\partial a_i} p_k(r \mid \underline{a})\, dr = \frac{\partial}{\partial a_i} \int p_k(r \mid \underline{a})\, dr = \frac{\partial}{\partial a_i} \left[1\right] = 0 \quad . \end{aligned} \tag{A-3.26}$$

Substituting the above result into (A-3.25) gives

$$\begin{aligned} J_{ij} &= \sum_{k=1}^{N} E[Y_{i,k}\, Y_{j,k}] = \sum_{k=1}^{N} E\left[\frac{\partial}{\partial a_i} \ln p_k(R_k \mid a)\, \frac{\partial}{\partial a_j} \ln p_k(R_k \mid \underline{a})\right] \\ &= \sum_{k=1}^{N} J_{ij}^{(k)} \quad , \end{aligned} \tag{A-3.27}$$

where $J^{(k)}$ is defined to be the J-matrix for estimating $\underline{a}$ based on R_k. This shows that $J = \sum_{k=1}^{N} J^{(k)}$, which is the desired result (A-3.3).

A.4 ADDITIVE GAUSSIAN NOISE

A.4.1 Environment 1

Suppose

$$\underline{R} = \underline{f}(\underline{A}) + \underline{N} \quad , \tag{A-4.1}$$

where $\underline{f}$ is a known vector-valued function and $\underline{N}^t = (N_1, \ldots, N_M)$ is a vector of Gaussian noise, independent of $\underline{A}$, with zero means and covariance matrix C_N,

$$C_N \triangleq E[\underline{N}\underline{N}^t] \, , \; [C_N]_{ij} = E[N_i N_j] \quad . \tag{A-4.2}$$

In this case

$$\begin{aligned} p_{\underline{R}|\underline{A}}(\underline{r} \mid \underline{a}) &= p_{\underline{N}}(\underline{r} - \underline{f}(\underline{a})) \\ &= \frac{1}{(2\pi)^{M/2}\sqrt{\det(C_N)}} \exp\left\{-\frac{1}{2}(\underline{r} - \underline{f}(\underline{a}))^t C_N^{-1}(\underline{r} - \underline{f}(\underline{a}))\right\} \quad . \end{aligned} \tag{A-4.3}$$

Substituting the above into (A-1.6a) and simplifying yields

$$J = \Gamma C_N^{-1} \Gamma^t \quad , \tag{A-4.4}$$

where Γ is the $L \times M$ Jacobean matrix defined by

$$\Gamma_{ij} = \frac{\partial}{\partial a_i} f_j(\underline{a}) \quad . \tag{A-4.5}$$

If $\underline{f}(\underline{a})$ is linear; i.e., if

$$\underline{R} = F\underline{a} + \underline{N} \quad , \tag{A-4.6}$$

where F is an $L \times M$ matrix, then $\Gamma = F^t$ and so

$$J = F^t C_N^{-1} F \quad . \tag{A-4.7}$$

Note that, just as in the linear scalar case (A-2.11), J does not depend on $\underline{a}$. Also, as in the linear scalar case, the ML estimator is unbiased and achieves the Cramer-Rao bound. Consequently, the Cramer-Rao bound is tight. Specifically, the ML estimator is

$$\hat{\underline{A}}(\underline{r}) = [F^t C_N^{-1} F]^{-1} F^t C_N^{-1} \underline{r} \tag{A-4.8}$$

and has error covariance matrix $C = J^{-1}$.

If $N_1, \ldots, N_M$ are independent, then (A-4.4) reduces to the answer given previously in (A-3.3) and (A-3.8).

If $N_1, \ldots, N_M$ are identical as well as independent, then (A-4.4) reduces to

$$J = \frac{1}{\sigma_N^2} \Gamma\Gamma^t \quad , \tag{A-4.9}$$

and (A-4.7) reduces to

$$J = \frac{1}{\sigma_N^2} F^t F \quad . \tag{A-4.10}$$

A.4.2 Environment 2

Let us again suppose that (A-4.1) holds; i.e., $\underline{R} = f(\underline{A}) + \underline{N}$, where $\underline{N}$ is Gaussian with zero means and covariance matrix C_N, and in addition, suppose that $\underline{A}^t = (A_1, \ldots, A_L)$ is Gaussian with zero means and covariance matrix C_A. Then

$$p_{\underline{A}}(\underline{a}) = \frac{1}{(2\pi)^{L/2}\sqrt{\det(C_A)}} \exp\left\{-\frac{1}{2}\underline{a}^t C_A^{-1}\underline{a}\right\} \quad , \qquad \text{(A-4.11)}$$

and direct substitution into (A-1.9) yields

$$K = C_A^{-1} \quad . \qquad \text{(A-4.12)}$$

Combining (A-4.7) and (A-4.12) gives

$$(\bar{J} + K) = \overline{\Gamma C_N^{-1}\Gamma} + C_A^{-1} \quad , \qquad \text{(A-4.13)}$$

where the overbar denotes expectation over $\underline{A}$.

In the linear case where $f(\underline{a}) = F\underline{a}$ and

$$\underline{R} = F\,\underline{A} + \underline{N} \quad , \qquad \text{(A-4.14)}$$

we obtain from (A-4.7) and (A-4.12), that

$$(\bar{J} + K) = F^t C_N^{-1} F + C_A^{-1} \quad . \qquad \text{(A-4.15)}$$

Once again in this linear Gaussian case it is known that the Cramer-Rao bound is tight in the sense that, for the best estimator $\hat{\underline{A}}^*(\underline{r}) = E[\underline{A} \mid \underline{r}]$, the covariance matrix of the errors equals $[\bar{J} + K]^{-1}$.

A.5 COMPLEX PARAMETERS AND MEASUREMENTS

In this section we discuss the situation in which either the parameters to be estimated or the measurements or both are complex vectors. We begin with some notation.

For reasons to be seen later, we will typically represent a complex vector such as $\underline{x} = (x_1, x_2, \ldots, x_M)$, wherein each component x_j is complex, by a vector $\underline{\mathcal{X}}$ with 2M real-valued components: specifically,

$$\underline{\mathcal{X}} = (x_{1r}, x_{1i}, x_{2r}, x_{2i}, \ldots, x_{Mr}, x_{Mi}) \quad , \tag{A-5.1}$$

where x_{jr} and x_{ji} are the real and imaginary parts, respectively, of x_j. Alternatively, we may write

$$\underline{\mathcal{X}} = (\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_M) \quad , \tag{A-5.2}$$

where $\underline{x}_j = (x_{jr}, x_{ji})$ is the vector representation of $x_j = (x_{jr} + ix_{ji})$. We will often write x_{ju} as generic notation to stand for either x_{jr} or x_{ji}. It will occasionally be convenient to flip back and forth between the two notations $\underline{x}$ and $\underline{\mathcal{X}}$, depending on what needs emphasis and which requires fewer symbols. For example, when magnitude or length are relevant, then the magnitude $|x_j|$ and the length $||\underline{x}_j||$ both denote $\sqrt{x_{jr}^2 + x_{ji}^2}$, but the former is simpler to write.

Note further that the "lengths" of the vectors $\underline{x}$ and $\underline{\mathcal{x}}$ are identical:

$$\begin{aligned} ||\underline{x}|| &\triangleq \left(\sum_{j=1}^{M} |x_j|^2\right)^{1/2} = \left(\sum_{j=1}^{M} ||\underline{x}_j||^2\right)^{1/2} = \left(\sum_{j=1}^{M} (x_{jr}^2 + x_{ji}^2)\right)^{1/2} \\ &= \left(\sum_{j=1}^{M} \sum_{u \in \{r,i\}} x_{ju}^2\right)^{1/2} \triangleq ||\underline{\mathcal{x}}|| \end{aligned} \tag{A-5.3}$$

A.5.1 Complex Measurements

When the measurements form a complex vector $R = (R_1,...,R_M)$, we will represent them as a real vector $\underline{R} = (R_{1r},R_{1i},...,R_{Mi})$. For one thing this allows there to be a well-defined probability density $p(\underline{\mathcal{x}} \mid \underline{a})$.

A.5.2 Estimating Complex Parameters

When the parameters to be estimated form a complex vector $A = A_1,...,A_L)$, we will separately estimate the real and imaginary components of each A_j. That is we will represent A by the vector $\underline{A} = (A_{1r},A_{1i},A_{2r},...,A_{Li})$ with real components and produce an estimate $\hat{\underline{A}} = (\hat{A}_{1r},\hat{A}_{1i},..., \hat{A}_{Li}) = (\hat{A}_1, \hat{A}_2,...,\hat{A}_L)$. If so desired, $\hat{\underline{A}}$ can be converted to a complex estimate $\hat{A} = (\hat{A}_1,...,\hat{A}_L)$ for A. The mean squared errors in $\hat{A}$ are related to those in $\hat{\underline{A}}$ by

$$|A_j - \hat{A}_j|^2 = (A_{jr} - \hat{A}_{j,r})^2 + (A_{ji} - \hat{A}_{ji})^2 = ||A_j - \hat{A}_j||^2 \quad . \tag{A-5.4}$$

Note that the Cramer-Rao bound requires differentiation of a certain function with respect to the parameters. This makes sense only if the parameters are

real and provides further motivation for replacing A by $\underset{\sim}{A}$. Specifically, the Cramer-Rao bound for estimating $\underset{\sim}{A}$ from a real vector R is, in our notation,

Environment 1:

$$E\left[\left(\hat{A}_{ku}-a_{ku}\right)^2\right] \geq [J]^{-1}_{ku,ku} \quad , \; k = 1,...,M, \; u \in \{r,i\} \quad , \tag{A-5.5}$$

where

$$J_{ku,lv} = E_R\left[\frac{\partial}{a_{ku}} \ln p(\underset{\sim}{r} \mid \underset{\sim}{a}) \frac{\partial}{\partial a_{lv}} \ln p(\underset{\sim}{r} \mid \underset{\sim}{a})\right] \quad , \tag{A-5.6}$$

Environment 2:

$$E[(\hat{A}_{ku}-A_{ku})^2] \geq [\bar{J}+K]^{-1}_{ku,ku} \quad , k = 1,...,M, \; u \in \{v,i\} \quad , \tag{A-5.7}$$

where

$$\bar{J}_{ku,lv} = E_{\underset{\sim}{A}}\left[J_{ku,lv}\right] \tag{A-5.8}$$

$$K_{ku,lv} = E_{\underset{\sim}{A}}\left[\frac{\partial}{\partial a_{ku}} \ln p(\underset{\sim}{a}) \frac{\partial}{\partial a_{lv}} \ln p(\underset{\sim}{a})\right] \quad . \tag{A-5.9}$$

A.5.3 Independent Measurements

Let us consider the commonly occurring situation in which the components of the complex measurement vector $R = (R_1,...,R_M)$ are conditionally independent given the real parameter vector $A = (A_1,...,A_L)$. With R represented by the real vector $\underset{\sim}{R} = (R_{1r},R_{1i},...,R_{Mi})$, the conditional independence means

$$p_{\underset{\sim}{R}|\underset{\sim}{A}}(\underset{\sim}{r}\,|\,\underset{\sim}{a}) = \prod_{k=1}^{M} p_k(r_k\,|\,\underset{\sim}{a}) \quad , \tag{A-5.10}$$

where $p_k(r_k\,|\,\underset{\sim}{a})$ is shorthand for $p_{R_k|\underset{\sim}{A}}(r_k\,|\,\underset{\sim}{a})$. This happens, for example, if

$$R_k = f_k(A,N_k) \quad , \quad k=1,\ldots,M \quad , \tag{A-5.11}$$

where $f_1,\ldots,f_M$ are known complex-valued functions and $N_1,\ldots,N_M$ are complex-valued random variables that are independent of each other and of A. The conditional independence considered here is but a slight generalization of that considered in Section A.3.1.

Arguments similar to those used in Section A.3.3 show that the J matrix of the Cramer-Rao lower bound to the MSE when estimating $\underset{\sim}{A}$ based on $\underset{\sim}{R}$ reduces to

$$J = \sum_{k=1}^{M} J^{(k)} \quad , \tag{A-5.12}$$

where $J^{(k)}$ is the J-matrix for estimating $\underset{\sim}{a}$ based on $R_k = (R_{kr},R_{ki})$; that is,

$$J_{ij}^{(k)} = E_{R_k}\left[\frac{\partial}{\partial a_i}\ln p_k(r_k\,|\,\underset{\sim}{a})\,\frac{\partial}{\partial a_j}\ln p_k(r_k\,|\,\underset{\sim}{a})\right] \quad . \tag{A-5.13}$$

Equation (A-5.2) shows that the effect of each complex measurement is additive.

A.5.4 One-for-one Independent Measurements

Let us now consider the situation in which there are M complex parameters $\underset{\sim}{A} = (A_1,\ldots,A_M)$, M complex measurements $\underset{\sim}{R} = (R_1,\ldots,R_M)$, and each measurement R_k depends only on A_k in the sense that

$$p_{\underset{\sim}{R}\mid\underset{\sim}{A}}(\underset{\sim}{r}\mid\underset{\sim}{a}) = \prod_{k=1}^{M} p_k(r_k \mid a_k) \tag{A-5.14a}$$

or equivalently,

$$p(r_k \mid \underset{\sim}{a}) = p_k(r_k \mid a_k) \quad , \tag{A-5.14b}$$

where $p_k(r_k \mid a_k) \triangleq p_{R_k \mid A_k}(r_k \mid a_k) = p_{R_k \mid \underset{\sim}{A}}(r_k \mid \underset{\sim}{a})$. This happens, for example, if

$$R_k = f_k(A_k, N_k), \quad k = 1,\ldots,M \quad , \tag{A-5.15}$$

where $f_1,\ldots,f_M$ are known complex-valued function and $N_1,\ldots,N_M$ are complex-valued random vectors that are independent of each other and of $\underset{\sim}{A}$.

This case is also a special case of that considered in Section A.5.3. Restating the result given there for the present case gives

$$J = \sum_{k=1}^{M} J^{(k)} \quad , \tag{A-5.16}$$

where $J^{(k)}$ is the J-matrix for estimating $\underset{\sim}{a}$ based on $\underline{R}_k = (R_{kr}, R_{ki})$. That is,

$$J^{(k)}_{mu,nv} = E_{\underline{R}_k}\left[\frac{\partial}{\partial a_{mu}} \ln p_k(r_k \mid \underset{\sim}{a}) \frac{\partial}{\partial a_{nv}} \ln p_k(r_k \mid \underset{\sim}{a})\right] , \qquad m,n \in \{1,\ldots,M\} \; , \; u,v \in \{r,i\} \quad . \tag{A-5.17}$$

Substituting, (A-5.14b) into the above gives $J^{(k)}_{mu,nv} = 0$ unless $m = n = k$. Hence J is tri-diagonal. Specifically,

$$J = \begin{bmatrix} J_{(1)} & & & \\ & J_{(2)} & & 0 \\ & & \ddots & \\ 0 & & & J_{(M)} \end{bmatrix} \tag{A-5.18}$$

where $J_{(k)}$ is the 2×2 matrix

$$\left[J_{(k)}\right]_{uv} = E_{R_k}\left[\frac{\partial}{\partial a_{ku}} \ln p(\underline{r}_k \mid \underline{a}_k) \frac{\partial}{\partial a_{kv}} \ln p(\underline{r}_k \mid \underline{a}_k)\right] \quad , u,v \in \{r,i\} \quad . \tag{A-5.19}$$

Here, $J_{(k)}$ is the J-matrix in the Cramer-Rao lower bound for estimating $\underline{a}_k$ based on R_k.

A.6 CRAMER-RAO BOUNDS IN THE PRESENCE OF AMBIGUITY

Consider the additive noise estimation problem where

$$R = f(A) + N \quad . \tag{A-6.1}$$

When f is not one-to-one, we will say there is *ambiguity* in the function f and the measurement R. In such cases, one cannot expect any estimator to perform well. Unfortunately, however, the Cramer-Rao bound may not reflect this, i.e, it may give a very low MSE even though the actual MSE is rather large. We illustrate this phenomenon in the scalar case and show that it is not a problem in the linear Gaussian case.

Example 1: Scalar Parameters

One may see from (A-2.8) and, for the Gaussian case, (A-2.9) that the CRLB will be the same for any function $\tilde{f}$ whose derivative has the same magnitude as that of f. To get a clearer picture, consider the situation where f has a continuous second derivative but is not one-to-one. Then f has an interval where it is constant ($f' = 0$) and/or a pair of intervals I_+, I_- such that f increases on I_+ ($f' > 0$) and f decreases through the same range on I_- ($f' < 0$). The CRLB will be made large (and appropriately so) by the intervals where $f' = 0$. On the other hand, it will not be affected by the existence of pairs of intervals where f increases and decreases, respectively, through the same range, i.e., it will be as small as the CRLB for the nonambiguous function

$$\hat{f}(a) \triangleq \int_{-\infty}^{a} |f'(x)|\, dx \quad , \tag{A-6.2}$$

for which one expects there are estimators with much lower MSE than for f. Thus, we conclude that in the presence of ambiguity, the CRLB can only be expected to give a reasonable lower bound to the MSE for the estimation problem involving $\hat{f}$.

As a concrete example, compare the estimation of A based on either but not both of the following two measurements

$$R_1 = f_1(A) + N \ , \ R_2 = f_2(A) + N \ , \tag{A-6.3}$$

where $f_1(a) = a^3$, $f_2(a) = |a^3|$, A and N are Gaussian, independent and

$\eta(0, \sigma_A^2)$, $\eta(0, \sigma_N^2)$, respectively. There is obvious ambiguity in the R_2 measurement. For either measurement, the CRLB gives

$$E[(\hat{A} - A)^2] \geq \frac{1}{27\sigma_A^4/\sigma_N^2 + 1/\sigma_A^2} \tag{A-6.4}$$

This is a reasonable bound for estimation based on R_1. (It tends to 0 as $\sigma_N^2 \to 0$; it tends to σ_A^2 as $\sigma_N^2 \to \infty$.) It is not reasonable for estimation based on R_2. Indeed if $\sigma_N^2 = 0$, then the best estimate for A based on R_2 is $\hat{A} = E[A \mid R_2] = 0$ with MSE $= \sigma_A^2$, whereas the CRLB lower bound is zero.

Finally, consider estimation based on R_2, when the density of A is replaced by

$$\tilde{p}(a) = \begin{cases} \frac{2}{\sqrt{2\pi}\sigma_A} \exp\{-\frac{a^2}{2\sigma_A^2}\}, & a \geq 0 \\ 0, & a < 0 . \end{cases} \tag{A-6.5}$$

Now the increased a priori information removes all ambiguity (at least with probability 1), and the CRLB (which is the same as before) is a reasonable bound to MSE based on R_2. Thus, we see that additional a priori information can modify the problem so the CRLB is useful.

Example 2: Linear IID Gaussian Case

Here,

$$\underline{R} = F\underline{A} + \underline{N} \quad , \tag{A-6.6}$$

where $\underline{R} = (R_1, \ldots, R_M)^t$, $\underline{A} = (A_1, \ldots, A_L)^t$, $\underline{N} = (N_1, \ldots, N_M)^t$, F is an $M \times L$ matrix, and $\underline{A}$ and $\underline{N}$ are IID Gaussian $\eta(0, \sigma_A^2)$ and $\eta(0, \sigma_N^2)$, respectively and independent of each other. As mentioned in Section A.4 the CRLB bound is tight and gives

$$E[(A_i - \hat{A}_i)^2] = \left[\frac{F^t F}{\sigma_N^2} + \frac{I}{\sigma_A^2}\right]^{-1}_{ii} \tag{A-6.7}$$

If F has rank less than L (for example, if $M < L$), there will be ambiguity in $\underline{R}$. Nevertheless, the CRLB is tight.

As a concrete example, suppose

$$R = A_1 + A_2 + N \quad , \tag{A-6.8}$$

so that $L = 2$, $M = 1$, $F = [1\ 1]$. Then

$$\bar{J}+K = \frac{1}{\sigma_N^2}\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix} + \frac{1}{\sigma_A^2}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \quad , \tag{A-6.9}$$

$$[\bar{J}+K]^{-1} = \frac{\sigma_A^2\,\sigma_N^2}{\sigma_N^2 + 2\sigma_A^2}\begin{bmatrix}\dfrac{\sigma_N^2+\sigma_A^2}{\sigma_A^2\,\sigma_N^2} & -\dfrac{1}{\sigma_N^2}\\ -\dfrac{1}{\sigma_N^2} & \dfrac{\sigma_N^2+\sigma_A^2}{\sigma_A^2\,\sigma_N^2}\end{bmatrix} \tag{A-6.10}$$

and it is easy to show that

$$E[(A_1 - \hat{A}_1)^2] = E[(A_2 - \hat{A}_2)^2] = \frac{\sigma_A^2(\sigma_N^2 + \sigma_A^2)}{\sigma_N^2 + 2\sigma_A^2} \quad . \tag{A-6.11}$$

Note that as $\sigma_N^2 \to 0$, MSE $\to \sigma_A^2/2$, and as $\sigma_N^2 \to \infty$, MSE $\to \sigma_A^2$.

Appendix B

A GENERAL APPROACH TO LOWER BOUNDS TO PHASE RETRIEVAL ERROR

Introduction to the Estimation Problem

The problem is to estimate a complex-valued image g from measurements of its Fourier transform in the presence of additive noise and random phase. The goal of this section is to compute lower bounds to the least possible mean squared error in such estimates. Specifically, the image is a complex-valued $m \times m$ array

$$g = [g_p : p = (p_1, p_2),\ p_1 = 0, 1, \ldots, m-1,\ p_2 = 0, 1, \ldots, m-1]$$

For conciseness let $M = \{0, 1, \ldots, m-1\}$, and let $M^2 = M \times M$, so $g = [g_p : p \in M^2]$. The image g is weighted by the elements of a real valued non-negative weighting array $w = [w_p : p \in M^2]$, forming $f = [f_p : p \in M^2]$ according to

$$f'_k = w_p g_p \quad . \tag{B-1}$$

The special case in which the w_p's are 0 or 1 is of particular interest.

The measurements upon which the estimates of g are based form a complex-valued array $S = [S_k : k = (k_1, k_2) \in M^2]$, where

$$S_k = h\ (F'_k \exp\{i\Phi_k\}) + N_k \quad , \tag{B-2}$$

where $N = [N_k : k \in M^2]$ is an array of complex-valued additive random noise,

$\Phi = [\Phi_k : k \in M^2]$ is a real-valued array of random phases, $F = [F_k : k \in M^2]$ is the Fourier transform of f, and h is a known complex-valued function. We will shortly discuss each term in (B-2) in detail, but before doing so, we introduce the vector notation to be employed.

Vector Notation

It will be convenient to employ one-dimensional vector notation for the various arrays and matrix notation for the various linear transformations. For example, a complex-valued $m \times m$ array such as $f = [f_p : p \in M^2]$ will be represented as a real-valued $2m^2$-dimensional column vector $\underset{\sim}{f}$, whose components are the real and imaginary parts of the components of f, ordered as shown below:

$$\begin{aligned} \underset{\sim}{f} &= (f_{00r}, f_{00i}, f_{01r}, f_{01i}, \ldots, f_{m-1,m-1,i})^t \\ &= (\mathcal{f}_{00}, \mathcal{f}_{01}, \ldots, \mathcal{f}_{m-1,m-1})^t \quad , \end{aligned} \tag{B-3}$$

where f_{pr} and f_{pi} denote, respectively, the real and imaginary parts of f_p, and where $\mathcal{f}_p \triangleq (f_{pr}, f_{pi})$. By way of comparison note that $f_p = (f_{pr} + i f_{pi})$. Note also that the complex number f_p and the two-component vector $\mathcal{f}_p$, consisting of its real and imaginary parts, represent the same object. It will be convenient to flip back and forth between the two notations, depending on what needs emphasis and what notation requires fewer symbols. For example, the magnitude

$|f_p|$ and the length $||\mathcal{L}_p||$ give the same value, but the former is simpler to write and so will almost always be used. We use similar notation for other arrays, such as g, F and S.

The Weighting Array

The effect of the weighting array w may be characterized as a matrix multiplication:

$$\underset{\sim}{f} = W\underset{\sim}{g} \tag{B-4}$$

where $W = [W_{pu,qv}: p,q \in M^2;\ u,v \in \{r,i\}]$ is the $2m^2 \times 2m^2$ diagonal matrix with

$$W = \begin{bmatrix} w_{00} & & & & 0 \\ & w_{00} & & & \\ & & w_{01} & & \\ & & & w_{01} & \\ 0 & & & & \ddots \end{bmatrix} \tag{B-5}$$

i.e.,

$$W_{pu,qv} = \begin{cases} w_p\ , & p = q,\ u = v \\ 0\ , & p \neq q \text{ or } u \neq v \end{cases} . \tag{B-6}$$

For brevity let $\mathbf{M}^2 \triangleq M^2 \times \{r,i\}$, so $W = [W_{s,t}: s,t \in \mathbf{M}^2]$.

The Fourier Transform

The Fourier transform relations are

$$F_k = \frac{1}{m} \sum_{p \in M^2} f_p \exp\{-i\frac{2\pi}{m}(k,p)\}, \quad k \in M^2 \tag{B-7a}$$

$$f_p = \frac{1}{m} \sum_{k \in M^2} F_k \exp\{i\frac{2\pi}{m}(k,p)\}, \quad p \in M^2, \tag{B-7b}$$

where $(k,p) \triangleq (k_1p_1+k_2p_2)$. Note that with the above definitions, Parseval's relation takes the form

$$\sum_{k \in M^2} |F_k|^2 = \sum_{p \in M^2} |f_p|^2 . \tag{B-8}$$

The Fourier transform may also be characterized as a matrix multiplication:

$$\underset{\sim}{F} = T\underset{\sim}{f} = TW\underset{\sim}{g} , \tag{B-9}$$

where $T = [T_{st} : s,t \in \mathbf{M}^2]$,

$$\underset{\sim}{T} = \begin{bmatrix} T_{00,00} & T_{00,01} & T_{00,02} & \cdots \\ T_{01,00} & T_{01,01} & & \\ T_{02,00} & & \ddots & \\ \vdots & & & \end{bmatrix} \tag{B-10}$$

$$T_{k,p} = \begin{bmatrix} T_{kr,pr} & T_{kr,pi} \\ T_{ki,pr} & T_{ki,pi} \end{bmatrix} = \begin{bmatrix} c(p,k) & s(p,k) \\ -s(p,k) & c(p,k) \end{bmatrix} \tag{B-11}$$

$$c(p,k) = \frac{1}{m} \cos\left[\frac{2\pi}{m}(p,k)\right] \qquad \textbf{(B-12a)}$$

$$s(p,k) = \frac{1}{m} \sin\left[\frac{2\pi}{m}(p,k)\right] \qquad \text{(B-12b)}$$

The matrix T is orthogonal; i.e., its rows (columns) are orthonormal, $T^t = T^{-1}$, and it follows that $||T\underset{\sim}{f}|| = ||\underset{\sim}{f}||$ for any $\underset{\sim}{f}$, which is simply Parseval's relation (B-5).

The Random Phase

The effect of the random phase array Φ can also be expressed as a matrix multiplication. Indeed the complete measurement can be expressed as

$$\underset{\sim}{S} = H(P\underset{\sim}{F}) + \underset{\sim}{N} = H(PTW\underset{\sim}{g}) + \underset{\sim}{N} \qquad \text{(B-13)}$$

where

$$H(\underset{\sim}{x}) = (\mathrm{Re}\{h(X_{00}\}, \mathrm{Im}\{h(X_{00})\}, \mathrm{Re}\{h(X_{01})\}, \ldots) \qquad \text{(B-14)}$$

and where $P = [P_{ku,lv}: k,l \in M^2;\ u,v \in \{r,i\}]$ is the real-valued $2m^2 \times 2m^2$ matrix

$$P = \begin{bmatrix} P_{00} & & & 0 \\ & P_{01} & & \\ & & \ddots & \\ 0 & & & P_{m-1,m-1} \end{bmatrix} \qquad \text{(B-15)}$$

$$P_k = \begin{bmatrix} P_{kr,kr} & P_{kr,ki} \\ P_{ki,kr} & P_{ki,ki} \end{bmatrix} = \begin{bmatrix} \cos \Phi_k & -\sin \Phi_k \\ \sin \Phi_k & \cos \Phi_k \end{bmatrix}, \quad k \in M^2 \qquad \text{(B-16)}$$

Note that P is orthogonal.

The Measurement Function

Examples of the measurement function include:

$$h(x) = ax \qquad \text{(B-17)}$$

$$h(x) = a \mid x \mid^2 \qquad \text{(B-18)}$$

$$h(x) = a \mid x \mid \quad , \qquad \text{(B-19)}$$

where "a" is a real-valued constant. Note that with (B-18) and (B-19), the resulting measurements $S_k = h(F_k \exp\{i\Phi_k\}) + N_k$ are not affected by the Φ_k's. Moreover, they are real-valued, so that the imaginary parts of S_k and N_k may be ignored.

Statistical Assumptions

We make the following statistical assumptions.

(1) The Additive Noise: The $2m^2$ components of $\underset{\sim}{N}$ are independent and identically distributed (IID). Consequently, the m^2 complex components of N are also IID. When it comes to making specific calculations, we will also assume that each component of $\underset{\sim}{N}$ is Gaussian with mean 0 and variance $\sigma_N^2/2$; i.e., they are $\eta(0,\sigma_N^2/2)$. This implies $E[\mid N_k \mid^2] = \sigma_N^2$.

(2) The Phase Noise: The m^2 components of Φ are IID with density $p_\Phi(\Phi)$. Specific calculations will later be made for the uniform density on $[0,2\pi]$. This models a complete loss of phase in the measurements S.

(3) The Image: The $2m^2$ components of $\underset{\sim}{g}$ are IID Gaussian, $\eta(0,\sigma_g^2/2)$. Consequently, the m^2 complex components of g are IID with $E[\,|\,g_k\,|^2] = \sigma_g^2$.

(4) The additive noise $\underset{\sim}{N}$, the phase notice Φ and the image $\underset{\sim}{g}$ are independent of each other.

These assumptions have two important consequences. First, the facts that $S_k = h(F_k\, e^{i\Phi_k}) + N_k$ and that $\underset{\sim}{N}$ and Φ are IID and independent of each other and of $\underset{\sim}{g}$ imply that each measurement S_k (equiv. $\underset{\sim}{S}_k$) depends only on the corresponding element F_k (equiv. $\underset{\sim}{F}_k$) of the transformed image; i.e., given F_k, the measurement S_k is conditionally independent of all other F_j's and S_j's. That is,

$$p(\underset{\sim}{S}\,|\,\underset{\sim}{F}) = \prod_{k\in M^2} p(\underset{\sim}{S}_k\,|\,\underset{\sim}{F}_k)\ . \tag{B-20}$$

Thus, in regard to estimating F, we have the situation of one-for-one independent complex measurements, as discussed in the section on the Cramer-Rao bound for complex parameters and measurements.

For future reference, the individual measurement statistics $p(\underset{\sim}{S}_k\,|\,\underset{\sim}{F}_k)$ may be derived as follows

$$p(\underset{\sim}{S}_k \mid \underset{\sim}{E}_k) = \int_0^{2\pi} p(\underset{\sim}{S}_k \mid \Phi_k, \underset{\sim}{E}_k)\, p(\Phi_k) d\Phi_k \quad , \tag{B-21}$$

where

$$p(\underset{\sim}{S}_k \mid \Phi_k, \underset{\sim}{E}_k) = p_{\underset{\sim}{N}_k}(\underset{\sim}{S}_k - h(F_k e^{i\Phi_k})) \tag{B-22}$$

The second consequence of these assumptions is that $\underset{\sim}{f}$ and $\underset{\sim}{F}$ are Gaussian, because they are linear combinations of Gaussian variables. Their means are zero and their covariance matrices are

$$C_f = E[\underset{\sim}{f}\,\underset{\sim}{f}^t] = WC_g W^t = \frac{\sigma_g^2}{2} W^2 \tag{B-23}$$

$$C_F = TC_f T^t = \frac{\sigma_g^2}{2} TW^2T^t \quad . \tag{B-24}$$

Since C_f is diagonal, the components of $\underset{\sim}{f}$ are independent with variances

$$\sigma^2_{f,u} = \frac{w_p^2 \sigma_g^2}{2} \quad . \tag{B-25}$$

They are not, however, identically distributed unless all weights w_p are identical.

We will next show that each component of $\underset{\sim}{F}$ is $N(0,\sigma_f^2/2)$, where

$$\sigma_F^2 = \sigma_g^2 \frac{1}{m^2} \sum_{p \in M^2} w_p^2 \quad , \tag{B-26}$$

and that $F_{k,r}$ and $F_{k,i}$ are independent, for any k. It is not true in general, that F_{ku} and F_{lv} are independent for $k \neq l$.

Using the independence of the components of $\underset{\sim}{f}$, we find that for any $u, v \in \{r, i\}$,

$$E[F_{ku} F_{lv}] = E\left[\sum_{pw} T_{ku,pw} f_{pw} \sum_{qz} T_{lv,qz} f_{qz}\right]$$
$$= \sum_{pw} E[f_{pw}^2] T_{ku,pw} T_{lv,pw} \tag{B-27}$$

Now if $k = l$ and $u = v$, we have

$$E[F_{ku}^2] = \sum_{pw} \frac{\sigma_f^2}{2} w_p^2 T_{ku,pw}^2$$
$$= \frac{\sigma_f^2}{2} \sum_p w_p^2 \frac{1}{m^2} \left[\cos^2\left[\frac{2\pi}{m}(p,k)\right] + \sin^2\left[\frac{2\pi}{m}(p,k)\right]\right]$$
$$= \frac{\sigma_f^2}{2m^2} \sum_p w_p^2 \quad , \tag{B-28}$$

and this establishes (B-26). Finally, if $k = l$, $u = r$ and $v = i$, then

$$E[F_{kr} F_{ki}] = \sum_p \frac{\sigma_f^2}{2} w_p^2 \left(T_{kr,pr} T_{ki,pr} + T_{kr,pi} T_{ki,pi}\right)$$
$$= \frac{\sigma_f^2}{2} \sum_p w_p^2 \frac{1}{m^2} \left(-c(p,k)s(p,k) + s(p,k)c(p,k)\right)$$
$$= 0 \quad , \tag{B-29}$$

and this implies that F_{kr} and F_{ki} are independent.

Bounds to Estimation Error

The problem is to estimate $\underset{\sim}{g} = [g_s : s \in \mathbf{M}^2]$ from $\underset{\sim}{S} = [S_t : t \in \mathbf{M}^2]$. We seek to lower bound the minimum possible mean squared errors. Let $\underset{\sim}{\hat{g}} = [\hat{g}_s : s \in \mathbf{M}^2]$ denote an estimate of $\underset{\sim}{g}$, let $\underset{\sim}{\tilde{g}} = (\underset{\sim}{g} - \underset{\sim}{\hat{g}})$ denote the resulting vector of errors, let $C_{\underset{\sim}{\tilde{g}}} = E[\underset{\sim}{\tilde{g}}\underset{\sim}{\tilde{g}}']$ denote the covariance matrix of $\underset{\sim}{\tilde{g}}$, let $\underset{\sim}{e}_{\tilde{g}} = [e_{\tilde{g},s} : s \in \mathbf{M}^2]$ denote the vector of mean squared errors (note: $e_{\tilde{g},s} = E[\tilde{g}_s^2]$ and $\underset{\sim}{e}_{\tilde{g}}$ is the diagonal of $C_{\tilde{g}}$), and let

$$e_{\tilde{g}} = \sum_{s \in \mathbf{M}^2} e_{\tilde{g},s} = tr(C_{\tilde{g}}) = E\left[\,||\underset{\sim}{g}||^2\right] = E[\underset{\sim}{\tilde{g}}'\underset{\sim}{\tilde{g}}] \qquad \text{(B-30)}$$

denote the total mean squared error. Actually, one is only interested in the errors in components of $\underset{\sim}{g}$ that are given non-zero weight. Correspondingly, let

$$e_{\tilde{g}}^{+} = \sum_{\substack{(p,u) \in \mathbf{M}^2: \\ e_{\tilde{g},p,u} > 0}} e_{\tilde{g},p,u} \qquad \text{(B-31)}$$

From now on we assume $\underset{\sim}{\hat{g}}$ is the optimum estimate of $\underset{\sim}{g}$, namely the one for which each component of $\underset{\sim}{e}_{\tilde{g}}$ is smallest, ($\underset{\sim}{\hat{g}} = E[\underset{\sim}{g} \mid \underset{\sim}{S}]$). We seek lower bounds to $\underset{\sim}{e}_{\tilde{g}}$, $e_{\tilde{g}}$, and $e_{\tilde{g}}^{+}$. Our approach is to first consider the problems of estimating $\underset{\sim}{f}$ and $\underset{\sim}{F}$ from $\underset{\sim}{s}$. We will show that the optimum estimates for $\underset{\sim}{f}$ and $\underset{\sim}{F}$ and the resulting mean squared errors are closely related to those for $\underset{\sim}{g}$. The

Cramer-Rao lower bound is most easily applied to the estimates of $\underset{\sim}{F}$, and such bounds may then be converted to provide bounds to the mean squared errors in the estimates of $\underset{\sim}{f}$ and $\underset{\sim}{g}$.

The fact that $\underset{\sim}{f} = W\underset{\sim}{g}$ suggests that $\hat{\underset{\sim}{f}} = W\hat{\underset{\sim}{g}}$ would be a good estimate for $\underset{\sim}{f}$. Indeed, it must be the optimum estimate for $\underset{\sim}{f}$, because if there were a better estimator $\hat{\hat{\underset{\sim}{f}}}$ for $\underset{\sim}{f}$, then the estimator

$$\hat{\hat{\underset{\sim}{g}}}_{pu} = \begin{cases} \frac{1}{w_p}\,\hat{\hat{f}}_{pu} & , \quad w_p > 0 \\ 0 & , \quad w_p = 0 \end{cases} \tag{B-32}$$

would be better for $\underset{\sim}{g}$ than $\hat{\underset{\sim}{g}}$, which would contradict the latter's optimality. We conclude from this argument that the optimal estimate $\hat{\underset{\sim}{f}}$ for $\underset{\sim}{f}$ is related to that of $\underset{\sim}{g}$ via

$$\hat{\underset{\sim}{f}} = W\hat{\underset{\sim}{g}} \tag{B-33}$$

$$\tilde{\underset{\sim}{f}} = W\,\tilde{\underset{\sim}{g}} \tag{B-34}$$

$$\underset{\sim}{e}_f = W^2\underset{\sim}{e}_g \tag{B-35}$$

$$C_{\tilde{f}} = WC_{\tilde{g}}W^t \tag{B-36}$$

$$e_f = \sum_{(p,u)\in M^2} w_p^2\, e_{g,p,u} \quad . \tag{B-37}$$

By the same argument the optimal estimate $\hat{\underset{\sim}{f}}$ for $\underset{\sim}{f}$ is related to the optimal estimate $\hat{\underset{\sim}{F}}$ for $\underset{\sim}{F}$ via

$$\hat{\underset{\sim}{F}} = T\hat{\underset{\sim}{f}} \tag{B-38}$$

$$\tilde{\underset{\sim}{F}} = T\tilde{\underset{\sim}{f}} \tag{B-39}$$

$$C_{\tilde{F}} = TC_{\tilde{f}}T^t \tag{B-40}$$

$$e_F = e_f \quad , \tag{B-41}$$

where the orthogonality of T was used to obtain (B-41).

We will now find the Cramer-Rao lower bound to the mean squared error in $\hat{\underset{\sim}{F}}$, and then use (B-33)-(B-41) to find lower bounds to the mean squared errors in $\hat{\underset{\sim}{f}}$ and $\hat{\underset{\sim}{g}}$. The Cramer-Rao lower bound applied to $\underset{\sim}{F}$ and $\hat{\underset{\sim}{F}}$ yields

$$e_{F,\theta} \geq [\bar{J}_F + K_F]^{-1}_{\theta\theta} \, , \quad \theta \in \mathbf{M}^2 \quad , \tag{B-42}$$

where $\bar{J}_F$ and K_F are the $2m^2 \times 2m^2$ matrices defined by

$$\bar{J}_F = E_{\underset{\sim}{F}}[J_F] \tag{B-43}$$

$$J_{F,\theta,t} = E_{\underset{\sim}{S}}\left\{ \frac{\partial}{\partial F_\theta} \ln p(\underset{\sim}{S} \mid \underset{\sim}{F}) \frac{\partial}{\partial F_t} \ln p(\underset{\sim}{S} \mid \underset{\sim}{F}) \right\} \tag{B-44}$$

$$K_{F,\theta,t} = E_{\underset{\sim}{F}}\left\{ \frac{\partial}{\partial F_\theta} \ln p(\underset{\sim}{F}) \frac{\partial}{\partial F_t} \ln p(\underset{\sim}{F}) \right\} \quad , \tag{B-45}$$

where $E_{\underset{\sim}{F}}$ and $E_{\underset{\sim}{S}}$ denote statistical averages over $\underset{\sim}{F}$ and $\underset{\sim}{S}$, respectively. The Cramer-Rao bound also gives the more complete result

$$C_{\underset{\sim}{F}} \geq [\bar{J}_F + K]^{-1} \quad , \tag{B-46}$$

where $A \geq B$ means that $A - B$ is a non-negative definite matrix. Note that since $A \geq B$ implies $A_{kk} \geq B_{kk}$, equation (B-46) implies (B-42).

Now from (B-40) and (B-46) we obtain a lower bound to the covariance matrix of the errors in $\hat{\underset{\sim}{f}}$:

$$\begin{aligned} C_{\underset{\sim}{f}} &= T^t C_{\underset{\sim}{F}} T \\ &\geq T^t [\bar{J}_F + K_F]^{-1} T \\ &= [T^t (\bar{J}_F + K_F) T]^{-1} \\ &= [T^t \bar{J}_F T + T^t K_F T]^{-1} \quad , \end{aligned} \tag{B-47}$$

where we have used the fact that $T^t A T$ is non-negative definite if A is, and consequently, $A \geq B$ implies $T^t A T \geq T^t B T$.

Finally, we obtain a lower bound to the mean squared errors in $\tilde{\underset{\sim}{q}}$ from (B-35) and (B-47):

$$e_{q,s} \geq \frac{e_{f,s}}{w_p^2} \geq \frac{[T^t \bar{J}_F T + T^t K_F T]^{-1}_{ss}}{w_p^2} \quad , \text{ if } s = (p,u) \text{ and } w_p > 0 \quad . \tag{B-48}$$

Clearly, $e_{q,s} = \sigma_q^2 / 2$ if $s = (p,u)$ and $w_p = 0$. It remains for us to compute $\bar{J}_F$ and K_F.

The matrix K_F:

Using the Gaussian nature of F, direct calculation (see also (A-4.12) in the discussion of the Cramer-Rao bound) gives

$$K_F = C_F^{-1} \ , \tag{B-49}$$

Thus the term involving K_F in (B-47) is

$$T^t K_F T = T^t C_F^{-1} T = C_f^{-1} = \frac{2}{\sigma_g^2} W^{-2} \ , \tag{B-50}$$

where we have used (B-23) and (B-24). Note in particular that $T^t K_F T$ is a diagonal matrix.

The matrix $\bar{J}_F$:

As mentioned earlier, the measurements in $S = [S_k = k \in M^2]$ are one-for-one and independent in regard to estimating $F = [F_k : k \in M^2]$. As discussed in Section A.5, this implies the matrix J_F has the form

$$J_F = \begin{bmatrix} J_{(00)} & & 0 \\ & J_{(01)} & \\ 0 & & \ddots \end{bmatrix} \tag{B-51}$$

where $J_{(k)}, k \in M^2$, is the 2×2 matrix in the Cramer-Rao bound for estimating E_k based on S_k. That is for any $u, v \in \{r, i\}$

$$[J_{(k)}]_{uv} = E\left[\frac{\partial}{\partial F_{ku}} \ln p(S_k \mid E_k) \frac{\partial}{\partial F_{kv}} \ln p(S_k \mid E_k)\right] . \qquad \text{(B-52)}$$

Since the joint distribution of F_k, S_k does not depend on k, the matrices $\bar{J}_{(k)} = E_{F_k}[J_{(k)}]$ are the same for all k, and we denote such by $\bar{J}_{(\cdot)}$. Then

$$\bar{J}_F = \begin{bmatrix} \bar{J}_{(\cdot)} & & 0 \\ & \bar{J}_{(\cdot)} & \\ 0 & & \ddots \end{bmatrix} . \qquad \text{(B-53)}$$

It remains only to find the 2×2 matrix $\bar{J}_{(\cdot)}$. This will be done for the three choices of h given in (B-17)-(B-19). We shall see that for each of these, $\bar{J}_{(\cdot)}$ has the form

$$\bar{J}_{(\cdot)} = \begin{bmatrix} d & 0 \\ 0 & d \end{bmatrix} \qquad \text{(B-54)}$$

where d depends on h and the statistics of $\underset{\sim}{N}, \Phi$ and $\underset{\sim}{F}$. Consequently, $\bar{J}_F$ is itself a diagonal matrix:

$$\bar{J}_F = \begin{bmatrix} d & & & 0 \\ & d & & \\ & & d & \\ 0 & & & \ddots \end{bmatrix} \qquad \text{(B-55)}$$

Assuming that $\bar{J}_F$ is diagonal we now substitute (B-50) and (B-55) into (B-47) and find

$$C\tilde{f} \geq \left[T^t \bar{J}_F T + T^t K_F T\right]^{-1}$$

$$= \left[dI + \frac{2}{\sigma_g^2} W^{-2}\right]^{-1}$$

$$= \begin{bmatrix} c_{00} & & & & & 0 \\ & c_{00} & & & & \\ & & c_{00} & & & \\ & & & c_{00} & & \\ & & & & \ddots & \\ 0 & & & & & \ddots \end{bmatrix} \tag{B-56}$$

where

$$c_p = \frac{1}{d + \frac{2}{\sigma_g^2 w_p^2}} \quad , p \in M^2 \ . \tag{B-57}$$

It follows that for any $s = (p, u)$

$$e_{f,s} = E\left[\tilde{f}_s^2\right] \geq c_p = \frac{1}{d + \frac{2}{\sigma_g^2 w_p^2}} \tag{B-58}$$

and

$$e_f = e_F \geq \sum_{p \in M^2} 2\, c_p = \sum_{p \in M^2} \frac{2}{d + \frac{2}{\sigma_g^2 w_p^2}} = \frac{\sigma_g^2 w_p^2}{1 + \frac{d\,\sigma_g^2 w_p^2}{2}} \ . \tag{B-59}$$

Finally, since $e_{g,s} = e_{f,s} / w_p^2$, we have

$$e_{g,s} \geq \frac{c_p}{w_p^2} = \frac{1}{d w_p^2 + 2/\sigma_g^2} \tag{B-60}$$

$$c_g \geq \sum_{p \in M^2} \frac{2c_p}{w_p^2} = \sum_{p \in M^2} \frac{2}{dw_p^2 + 2/\sigma_g^2} \quad . \tag{B-61}$$

and

$$c_g^+ \triangleq \sum_{\substack{(p,u) \in M^2 \\ w_p > 0}} c_{g,pu} \geq \sum_{\substack{p \in M^2 \\ w_p > 0}} \frac{2c_p}{w_p^2} = \sum_{\substack{p \in M^2 \\ w_p > 0}} \frac{2}{dw_p^2 + 2/\sigma_g^2} \quad . \tag{B-62}$$

In the special case where n pixels are weighted by $w_p = 1$ and the remaining pixels are weighted by $w_p = 0$,

$$c_g^+ = \frac{2n}{d + 2/\sigma_g^2} \quad . \tag{B-63}$$

The Three Special Cases

Let us now assume:

(i) The components of the additive noise $\underset{\sim}{N}$ are Gaussian $N(0,\sigma_N^2/2)$.

(ii) The random phase Φ is uniformly distributed on $[0,2\Pi]$.

For each of the three choices of the measurement function h given in (B-17)-(B-19), we will show that the matrix $\bar{J}_{(\cdot)}$ has the form given in (B-54) and we will find an expression for the paramater d.

Case 1: $h(x) = ax$

Recall that $\bar{J}_{(\cdot)}$ is the matrix associated with the Cramer-Rao lower bound for estimating $E_k = (F_{kr}, F_{ki})$ from $S_k = (S_{kr}, S_{ki})$ where

$$S_k = aF_k e^{j\Phi_k} + N_k \quad , \tag{B-64}$$

where (F_{kr}, F_{ki}) are independent $N(0, \sigma_F^2/2)$, $\sigma_F^2 = \sigma_g^2 m^{-2} \sum_p w_p^2$, where Φ_k is uniformly distributed on $[0, 2\pi]$, where (N_{kr}, N_{ki}) are independent $N(0, \sigma_N^2/2)$, and where E_k, Φ_k and N_k are mutually independent. To simplify notation, from now on we omit all subscripts k; e.g. we write $S = (S_r, S_i)$ instead of $S_k = (S_{kr}, S_{ki})$. Also recall that for $u, v \in \{r, i\}$

$$[J_{(\cdot)}]_{uv} \triangleq E_S \left[\frac{\partial}{\partial F_u} \ln p(S \mid E) \frac{\partial}{\partial F_v} \ln p(S \mid E) \right] . \tag{B-65}$$

Using the specific form of h and the densities of N and Φ in (B-21), (B-22) and integrating gives the measurement statistics

$$\begin{aligned} p(S \mid E) &= \int_0^{2\pi} \frac{1}{\pi\sigma_N^2} e^{-\frac{|S - aF e^{j\Phi}|^2}{\sigma_N^2}} \frac{1}{2\pi} d\Phi \\ &= \frac{1}{\pi\sigma_N^2} I_0 \left(\frac{a|S||F|}{\sigma_N^2/2} \right) e^{-\frac{|S|^2 + a^2|F|^2}{\sigma_N^2}} , \end{aligned} \tag{B-66}$$

where $I_0(\cdot)$ denotes the zeroth-order modified Bessel function of the first kind; i.e.

$$I_0(x) = \frac{1}{2\pi} \int_0^{2\pi} e^{x \cos\lambda} d\lambda \quad . \tag{B-67}$$

For future reference we note that since $p(S \mid E)$ depends on E only through $|F|$, we have

$$p(\underline{S} \mid |F|) = p(\underline{S} \mid \underline{E}) \quad . \tag{B-68}$$

Also for future reference we now show that the conditional density of the magnitude $|S|$ given $\underline{E}$ (or $|F|$) is Rician:

$$\begin{aligned} p_{|S|}(s \mid \underline{E}) &= \frac{d}{ds} Pr(|S| \leq s \mid \underline{E}) \\ &= \frac{d}{ds} \int_{|\underline{S}| \leq s} \frac{1}{\pi\sigma_N^2} I_0\left(\frac{a|S||F|}{\sigma_N^2/2}\right) e^{-\frac{|S|^2 + a^2|F|^2}{\sigma_N^2}} d\underline{S} \\ &= \frac{d}{ds} \int_0^s \frac{1}{\pi\sigma_N^2} I_0\left(\frac{ar|F|}{\sigma_N^2/2}\right) e^{-\frac{r^2 + a^2|F|^2}{\sigma_N^2}} 2\pi r \, dr \\ &= \frac{s}{\sigma_N^2/2} I_0\left(\frac{as|F|}{\sigma_N^2/2}\right) e^{-\frac{s^2 + a^2|F|^2}{\sigma_N^2}} , \end{aligned} \tag{B-69}$$

which has the form of a Rician density. Since $p(|S| \mid \underline{E})$ depends on $\underline{E}$ only through $|F|$, we also have

$$p(|S| \mid |F|) = p(|S| \mid \underline{E}) = \frac{|S|}{\sigma_N^2/2} I_0\left(\frac{a|S||F|}{\sigma_N^2/2}\right) e^{-\frac{|S|^2 + a^2|F|^2}{\sigma_N^2}} \tag{B-70}$$

To compute $J_{(\cdot)}$, we find using (B-66)

$$\ln p(\underline{S} \mid \underline{E}) = \ln I_0\left(\frac{a|S||F|}{\sigma_N^2/2}\right) + \ln \frac{1}{\pi\sigma_N^2} - \frac{|S|^2 + a^2|F|^2}{\sigma_N^2} \tag{B-71}$$

$$\frac{\partial}{\partial F_u} \ln p(\underline{S} \mid \underline{E}) = \frac{I'_0\left(\frac{a\,|S|\,|F|}{\sigma_N^2/2}\right)}{I_0\left(\frac{a\,|S|\,|F|}{\sigma_N^2/2}\right)} \frac{a\,|S|}{\sigma_N^2/2} \frac{\partial |F|}{\partial F_u} - \frac{a^2}{\sigma_N^2} \frac{\partial |F|^2}{\partial F_u}$$

$$- \frac{2a}{\sigma_N^2} F_u \left[G_0\left(\frac{a\,|S|\,|F|}{\sigma_N^2/2}\right) \frac{|S|}{|F|} - a \right] \tag{B-72}$$

where $I'_0(z)$ denotes the derivative of $I_0(z)$, where

$$G_0(z) \triangleq \frac{I'_0(z)}{I_0(z)} \tag{B-73}$$

and where we used the facts that

$$\frac{\partial |F|}{\partial F_u} = \frac{F_u}{|F|}, \quad \frac{\partial |F|^2}{\partial F_u} = 2F_u . \tag{B-74}$$

Then substituting (B-72) into (B-65) yields

$$[J_{(\cdot)}]_{uv} = \frac{4a^2}{\sigma_N^4} F_u F_v E_{\underline{S}} \left[\left(G_0\left(\frac{a\,|S|\,|F|}{\sigma_N^2/2}\right) \frac{|S|}{|F|} - a \right)^2 \right] . \tag{B-75}$$

We next show that the term inside the inner parentheses has expected value zero. Specifically using (B-66)

$$E_S\left[G_o\left(\frac{a|S||F|}{\sigma_N^2/2}\right)\frac{|S|}{|F|}\right] = \int_0^\infty G_o\left(\frac{a|S||F|}{\sigma_N^2/2}\right)\frac{|S|}{|F|}\, p(|S|\,||F|)d|S|$$

$$= \int_0^\infty \frac{I'_o\left(\frac{a|S||F|}{\sigma_N^2/2}\right)}{I_o\left(\frac{a|S||F|}{\sigma_N^2/2}\right)}\frac{|S|}{|F|}\,\frac{|S|}{\sigma_N^2/2}\, I_o\left(\frac{a|S||F|}{\sigma_N^2/2}\right) e^{-\frac{|S|^2+a^2|F|^2}{\sigma_N^2}}\, d|S|$$

$$= \frac{2}{\sigma_N^2|F|}\, e^{-\frac{a^2|F|^2}{\sigma_N^2}} \int_0^\infty I'_o\left(\frac{a|S||F|}{\sigma_N^2}\right)|S|^2 e^{-\frac{|S|^2}{\sigma_N^2}}\, d|S|$$

$$= \frac{2}{\sigma_N^2|F|}\, e^{-\frac{a^2|F|}{\sigma_N^2}}\,\frac{|F|a\sigma_N^2}{2}\, e^{\frac{|F|^2a^2}{\sigma_N^2}} = a\ , \tag{B-76}$$

where we used the integral formula

$$\int_0^\infty I'_o(Ax)x^2e^{-x^2B}\,dx = \frac{A}{4B^2}\, e^{\frac{A^2}{4B}}\ , \tag{B-77}$$

with

$$A = \frac{2|F|a}{\sigma_N^2}\ ,\ B = \frac{1}{\sigma_N^2}\ . \tag{B-78}$$

Using (B-76) in (B-75) gives

$$\left[J_{(\cdot)}\right]_{uv} = \frac{4a^2}{\sigma_N^4}\, F_u F_v (H(|F|)-a^2)\ , \tag{B-79}$$

where

$$H(|F|) \triangleq E_S\left[\left[G_o\left(\frac{a|S||F|}{\sigma_N^2}\right)\frac{|S|}{|F|}\right]^2\right]. \qquad \text{(B-80)}$$

Thus, $J_{(o)}$ has the form

$$J_{(o)} = \frac{4a^2}{\sigma_N^4}(H(|F|)-a^2)\begin{bmatrix} F_r^2 & F_rF_i \\ F_rF_i & F_i^2 \end{bmatrix}. \qquad \text{(B-81)}$$

The next step is to find $\bar{J}_{(o)}$, the expectation of $J_{(o)}$ over $|F|$. To this end, observe that

$$E_E\left[(H(|F|)-a^2)F_rF_i\right] = E_E\left[H(|F|)F_rF_i\right] - a^2E_E\left[F_rF_i\right]. \qquad \text{(B-82)}$$

The second term is zero because F_r and F_i are uncorrelated with zero means. The first term is also zero because F_r and F_i are also conditionally uncorrelated with conditionally zero means when given any value of $|F|$. Thus

$$E_E\left[(H(|F|)-a^2)F_rF_i\right] = 0 \qquad \text{(B-83)}$$

Next, observe that

$$E_E\left[(H(|F|)-a^2)F_r^2\right] = E_E\left[((H(|F|)-a^2)F_i^2\right]. \qquad \text{(B-84)}$$

Finally, using (B-81), (B-83) and (B-84), we obtain

$$\bar{J}_{(o)} = \begin{bmatrix} d & 0 \\ 0 & d \end{bmatrix}, \qquad \text{(B-85)}$$

where

$$d = \frac{2a^2}{\sigma_N^2}\left(b - \frac{a^2\sigma_F^2}{\sigma_N^2}\right) \tag{B-86}$$

and

$$\begin{aligned} b &= \frac{2}{\sigma_N^2} E_E\left[H(|F|)F_r^2\right] \\ &= \frac{2}{\sigma_N^2} E_{E,S}\left[\left(G_o\left(\frac{a|S||F|}{\sigma_N^2}\right)\frac{|S|}{|F|}F_r\right)^2\right]. \end{aligned} \tag{B-87}$$

We have not been able to find a closed form expression for b. However, using the fact $F_r = |F|\cos\theta$, where $|F|$ and θ are independent random variables with θ uniformly distributed on $[0,2\pi]$ and with $|F|$ Rayleigh distributed with density

$$p(|F|) = \frac{|F|}{\sigma_F^2/2} e^{-\frac{|F|^2}{\sigma_F^2}}, \tag{B-88}$$

and using (B-87) and (B-88) and simplifying a bit gives

$$b = \frac{2}{\sigma_N^2}\int_0^{\infty}\int_0^{\infty}\int_0^{2\pi} G_o^2\left(\frac{a|S||F|}{\sigma_N^2}\right)\frac{|S|^2}{|F|^2}(|F|\cos\theta)^2 p(\theta)p(|F|)p(|S|\,||F|)\,d\theta\,d|F|\,d|S|$$

$$= \frac{\sigma_N^4}{\sigma_F^2 a^4}\int_0^{\infty}\int_0^{\infty}\frac{(I'_o(|S||F|))^2}{I_o(|S||F|)}|S|^3|F|\,e^{-|F|^2\left(\frac{1}{\sigma_F^2}+\frac{a^2}{\sigma_N^2}\right)-|S|^2\frac{\sigma_N^2}{4a^2}}\,d|F|\,d|S| \tag{B-89}$$

Notice that b depends only on σ_N/a and σ_F.

Case 2: $h(x) = a|x|^2$

This case was previously analyzed in Section 2. Here

$$S_k = a \mid F_k \mid^2 + N_k \ , \tag{B-90}$$

where (F_{kr}, F_{ki}) are independent $n(0,\sigma_F^2/2)$, $\sigma_F^2 = \sigma_g^2 m^{-2} \sum_p w_p^2$, where N_k is real-valued $\eta(0,\sigma_N^2)$ and where E_k and N_k are mutually independent.

Proceeding as in Case 1, we drop the subscript k and find the measurement statistics

$$\begin{aligned} p(S \mid E) &= p_N(S - a \mid F \mid^2) \\ &= \frac{1}{\sqrt{\pi\sigma_N^2}} \, e^{-\frac{(S - a \mid F \mid^2)^2}{\sigma_N^2}} \ . \end{aligned} \tag{B-91}$$

Then

$$\ln p(S \mid E) = \ln \frac{1}{\sqrt{\pi\sigma_N^2}} - \frac{(S - a \mid F \mid^2)^2}{\sigma_N^2} \tag{B-92}$$

$$\begin{aligned} \frac{\partial}{\partial F_u} \ln p(S \mid E) &= \frac{4a}{\sigma_N^2} F_u (S - a \mid F \mid^2) \\ &= \frac{4a}{\sigma_N^2} F_u N \end{aligned} \tag{B-93}$$

$$\begin{aligned} \left[J_{(\cdot)} \right]_{uv} &= E_S \left[\frac{\partial}{\partial F_u} \ln p(S \mid E) \frac{\partial}{\partial F_v} \ln p(S \mid E) \right] \\ &= \frac{16a^2}{\sigma_N^4} F_u F_v E[N^2] = \frac{8a^2}{\sigma_N^2} F_u F_v \end{aligned} \tag{B-94}$$

$$J_{(\bullet)} = \frac{8a^2}{\sigma_N^2} \begin{bmatrix} F_r^2 & F_r F_i \\ F_r F_i & F_r^2 \end{bmatrix} \tag{B-95}$$

Now, averaging over E gives

$$\left[\bar{J}_{(\bullet)}\right]_{uv} = \begin{cases} \dfrac{4a^2\sigma_\beta^2}{\sigma_N^2} & , \ u = v \\ 0 & , \ u \neq v \end{cases} \tag{B-96}$$

$$\bar{J}_{(\bullet)} = \begin{bmatrix} d & 0 \\ 0 & d \end{bmatrix} \tag{B-97}$$

where

$$d = \frac{4a^2\sigma_\beta^2}{\sigma_N^2} = \frac{4a^2\sigma_I^2}{m^2\sigma_N^2} \sum_p w_p^2 \ . \tag{B-98}$$

Notice that d is a kind of signal to noise ratio. Substitution into (B-60) gives a result in agreement with that in Section 2.

Case 3: $h(x) = a\,|\,x\,|$

Here

$$S_k = a\,|\,F_k\,| + N_k \quad , \tag{B-99}$$

where F_{kr}, F_{ki} are independent $\eta(0,\sigma_F^2/2)$, $\sigma_F^2 = \sigma_f^2 m^{-2}\sum_p w_p^2$, where N_k is real-valued $\eta(0,\sigma_N^2)$ and where E_k and N_k are mutually independent.

Proceeding as in Case 2, we drop the subscript k and find the measurement statistics

$$p(S\,|\,E) = p_N(S - a\,|\,F\,|) = \frac{1}{\sqrt{\pi\sigma_N^2}}\, e^{-\frac{(S-a\,|\,F\,|)^2}{\sigma_N^2}} \quad . \tag{B-100}$$

Then

$$\ln p(S\,|\,E) = \ln \frac{1}{\sqrt{\pi\sigma_N^2}} - \frac{(S-a\,|\,F\,|)^2}{\sigma_N^2} \tag{B-101}$$

$$\frac{\partial}{\partial F_u} \ln p(S\,|\,E) = \frac{2aF_u}{\sigma_N^2\,|\,F\,|}\,(S - a\,|\,F\,|)$$

$$= -\frac{2aF_u}{\sigma_N^2\,|\,F\,|}\,N \tag{B-102}$$

$$\left[J_{(\cdot)}\right]_{uv} = E_S\left[\frac{\partial}{\partial F_u}\ \ln p(S \mid \underline{E})\ \frac{\partial}{\partial F_v}\ \ln p(S \mid \underline{E})\right]$$

$$= \frac{4a^2}{\sigma_N^4}\ \frac{F_u F_v}{|F|^2}\ E[N^2]$$

$$= \frac{2a^2}{\sigma_N^2}\ \frac{F_u F_v}{|F|^2} \qquad \text{(B-103)}$$

Averaging over $\underline{E}$ gives

$$\left[\bar{J}_{(\cdot)}\right]_{uv} = \begin{cases} \dfrac{a^2}{\sigma_N^2} & , \ u = v \\ 0 & , \ u \neq v \end{cases} \qquad \text{(B-104)}$$

where we have used the fact that $E[F_u^2/\ |F|^2] = 1/2$. Hence

$$\bar{J}_{(\cdot)} = \begin{bmatrix} d & 0 \\ 0 & d \end{bmatrix} \qquad \text{(B-105)}$$

where

$$d = \frac{a^2}{\sigma_N^2} \ . \qquad \text{(B-106)}$$

Notice that here d does not depend at all on σ_F^2.

Appendix C

PHASE RETRIEVAL FOR DISCRETE FUNCTIONS WITH SUPPORT CONSTRAINTS: SUMMARY

Thomas R. Crimmins

Environmental Research Institute of Michigan

P. O. Box 8618

Ann Arbor, Michigan 48107-8618

Abstract

A phase retrieval uniqueness theorem for two-dimensional discrete functions is presented. Also, closed-form reconstruction algorithms and an algorithm for generating them are described.

published in:
Topical Meeting on Signal Recovery and Synthesis II, Technical Digest (Optical Society of America, Washington, D.C.), presented 2-4 April 1986, Honolulu, Hawaii, pp. 75-78.

Phase Retrieval for Discrete Functions with Support Constraints: Summary

Thomas R. Crimmins
Environmental Research Institute of Michigan
P. O. Box 8618, Ann Arbor, Michigan 48107-8618

1. Introduction

The phase retrieval problem, i.e., the problem of reconstructing a function from its Fourier modulus or, equivalently, from its autocorrelation function, arises in many fields, e.g., astronomy, wave-front sensing, X-ray crystallography, electron microscopy, particle scattering and pupil-function determination. Here we consider the case in which the object function is assumed to be defined on a two-dimensional discrete grid of sample points.

Bruck and Sodin [1] showed that uniqueness of phase retrieval in this case is equivalent to the irreducibility of a polynomial in two variables which is closely associated with the Fourier transform (z-transform) of the object function. Fiddy, Brames, and Dainty [2] used Eisenstein's irreducibility criterion to prove uniqueness for discrete object functions satisfying certain support constraints. Fienup [3] showed, among other things, that object functions satisfying certain different support constraints ("triangular objects") are uniquely defined by their autocorrelation functions among all other object functions satisfying the same support constraints. He also presented a closed-form algorithm for reconstructing such object functions from their autocorrelation functions.

A generalization of Fienup's results to a wider class of support constraints is presented here. Also, an algorithm for generating closed-form reconstruction algorithms is described. B. J. Brames recently obtained a result (unpublished) similar to the uniqueness theorem in Section 3.

Proofs will be omitted in this summary.

2. Masks

Let R^2 denote the Euclidean plane and let Z^2 denote the points in R^2 with integer coordinates. A finite subset of Z^2 is a mask if it contains at least three non-collinear points and its convex hull in R^2 has no parallel sides. Let M be a mask and let [M] denote its convex hull in R^2. Then [M] is a convex polygon (including its interior). A vertex v of [M] is opposite a side s of [M] if the line through v and parallel to s contains no points of [M] other than v. A vertex of [M] is a reference point of M if it is opposite some side of [M]. The set of all reference points of M will be denoted by R(M).

3. Uniqueness Theorem

Let f and f_1 be complex-valued functions on Z^2. The support of a function on Z^2 is the set of points at which the function is non-zero. Let S(f) and $S(f_1)$ denote the supports of f and f_1 and let r and r_1 be the autocorrelation functions of f and f_1, respectively. We have the following uniqueness theorem.

Theorem: If M is a mask, $R(M) \subseteq S(f) \subseteq M$, $S(f_1) \subseteq M$ and $r = r_1$, then there exists a complex number α of modulus 1 such that $f_1 = \alpha f$.

4. Reconstruction Algorithms

In this section closed-form algorithms for reconstructing a function from its autocorrelation function will be described.

Let S be the number of vertices of [M]. Let $v_0, \ldots, v_{S-1}$ be an ordering of the vertices going around [M] in the counter-clockwise direction and let $p_0, \ldots, p_{T-1}$ be a similar ordering of the reference points of M. It can be shown that R(M) contains an odd number of points so that T is odd. Let $K = (T-1)/2$ and let $q_n = p_{(nK) \bmod T}$ for $n = 0, \ldots, T-1$. Since K and T are relatively prime, the q_n are distinct and hence run through all the points of R(M). It can be shown that q_n and $q_{(n+1) \bmod T}$ have unique separation in M. That is, if $x, y \in M$ and $x - y = q_{(n+1) \bmod T} - q_n$ then $x = q_{(n+1) \bmod T}$ and $y = q_n$.

Let N be the number of points in M. A reconstruction algorithm for the mask M is an ordered pair, (q, m), where $q = (q_0, \ldots, q_{N-1})$ is an ordering of the points in M and $m = (m_T, \ldots, m_{N-1})$ is a sequence of integers satisfying the following conditions. The points $q_0, \ldots, q_{T-1}$ are as described above. For $n = T, \ldots, N-1$, the integers m_n satisfy the condition $0 \leq m_n \leq T-1$, and $M \cap (M + q_n - q_{m_n}) \subseteq \{q_0, \ldots, q_n\}$ and $M \cap (M - q_n + q_{m_n}) \subseteq \{q_0, \ldots, q_{n-1}\}$.

In the next section an algorithm for generating such reconstruction algorithms will be described.

In order to justify the above definition of reconstruction algorithms it will now be shown how they can be used to reconstruct a function from its autocorrelation function.

Let f be a complex-valued function on Z^2 satisfying $R(M) \subseteq S(f) \subseteq M$ and let r be its autocorrelation function. Since q_n and $q_{(n+1) \bmod T}$ have unique separation in M, it follows that $r(q_{(n+1) \bmod T} - q_n) = f(q_{(n+1) \bmod T}) \cdot f(q_n)^*$ where the * indicates complex conjugation. Therefore,

$$|f(q_0)|^2 = \frac{\prod_{n=0}^{K} r(q_{2n} - q_{(2n+1) \bmod T})}{\prod_{n=0}^{K-1} r(q_{2n+2} - q_{2n+1})} \quad . \tag{1}$$

Since f is defined by r only up to multiplication by a modulus 1 complex scalar, we may require that $f(q_0) > 0$. Then $f(q_0)$ is equal to the positive

square root of the right-hand side of Eq. 1. Now $f(q_n)$ can be computed for $n = 1, \ldots, T - 1$ from the formula $f(q_n) = r(q_n - q_{n-1})/f(q_{n-1})^*$. It can be shown that if (q, m) is a reconstruction algorithm then the following program will compute $f(q_n)$ for $n = T, \ldots, N - 1$.

Set $f(q_n) = 0$ for $n = T, \ldots, N - 1$ and set $n = T - 1$.

Step 1: If $n = N - 1$, stop. Otherwise $n \leftarrow n + 1$.

Step 2: $$f(q_n) = \left[r(q_n - q_{m_n}) - \sum_{k=0}^{n-1} f(q_k)\, f(q_k - q_n + q_{m_n})^* \right] / f(q_{m_n})^*.$$

Step 3: Go to Step 1.

5. Algorithm for Generating Reconstruction Algorithms

It will be assumed that we are given a sequence of all vertices $v_0, \ldots, v_{S-1}$ of [M] where M is a mask and the sequence is ordered in the counter-clockwise direction around [M].

For $n = 0, \ldots, S - 1$, let s_n be the side of [M] with endpoints v_n and $v_{(n+1) \bmod S}$. Let U be the linear operator on R^2 which rotates each vector in R^2 90° counter-clockwise.

First, the reference points $p_0, \ldots, p_{T-1}$ must be found. Note that every side of [M] has a vertex opposite it which is therefore a reference point. Of course, several sides may have the same vertex opposite them. Let $w_n = v_{(n+1) \bmod S} - v_n$ for $n = 0, \ldots, S - 1$. It can be shown that a vertex v_m is opposite a side s_n if and only if $\langle v_m, Uw_n \rangle \geq \langle v_k, Uw_n \rangle$ for $k = 0, \ldots, S - 1$, where $\langle , \rangle$ denotes the inner product on R^2. Thus, by taking each side in the order $s_0, \ldots, s_{S-1}$ all the reference points of M will be found and if they are numbered in the order in which they are found, $p_0, \ldots, p_{T-1}$, then the ordering will be in the counter-clockwise direction around [M].

As mentioned above T is odd. Let $K = (T - 1)/2$ and $q_n = p_{(nK) \bmod T}$, $n = 0, \ldots, T - 1$. Since each reference point, q_n, is a vertex of [M], there exists an integer k_n such that $0 \leq k_n \leq S - 1$ and $q_n = v_{k_n}$. For $n = 0, \ldots, T - 1$, define $y_n = Uw_{(k_n - 1) \bmod S} - Uw_{k_{(n+1) \bmod T}}$. Then it can be shown that for $x \in M$, $x \neq q_n$ and $x \neq q_{(n+1) \bmod T}$, $\langle q_n, y_n \rangle < \langle x, y_n \rangle < \langle q_{(n+1) \bmod T}, y_n \rangle$. (The uniqueness of separation of q_n and $q_{(n+1) \bmod T}$ mentioned in Section 4 follows from this double inequality.)

Now let $\alpha_n = q_n + q_{(n+1) \bmod T}$ and let $\beta_n = \alpha_n/2$ for $n = 0, \ldots, T-1$. Let $D = M \setminus R(M)$ (set difference) and let ϕ be the characteristic function of D as a subset of Z^2. That is, ϕ is the function on Z^2 which is 1 on D and 0 outside D. The set D contains $N - T$ points. For $x \in D$ and $0 \leq n \leq T-1$, let $h_n(x) = \langle x - \beta_n, y_n \rangle$. Now define T orderings of the points in D, $D = \{d_{n,0}, \ldots, d_{n,N-T-1}\}$, $n = 0, \ldots, T-1$, satisfying $|h_n(d_{n,k})| \geq |h_n(d_{n,k+1})|$ for $k = 0, \ldots, N-T-2$.

Set $n = T-1$ and $k = 0$ and enter the following loop.

Step 1. If $n = N-1$, stop. Otherwise define
$t = \min\{j : 0 \leq j \leq N-T-1 \text{ and } \phi(d_{k,j}) = 1\}$.

Step 2. If $\phi(\alpha_k - d_{k,t}) = 1$, go to Step 7.

Step 3. $n \leftarrow n + 1$.

Step 4. Define $q_n = d_{k,t}$.

Step 5. If $h_k(q_n) \geq 0$, define $m_n = k$. Otherwise define $m_n = (k+1) \bmod T$.

Step 6. $\phi(q_n) \leftarrow 0$.

Step 7. $k \leftarrow (k+1) \bmod T$ and go to Step 1.

It can be shown that the loop is not infinite and if $q = (q_0, \ldots, q_{N-1})$ and $m = (m_T, \ldots, m_{N-1})$ then (q, m) is a reconstruction algorithm.

6. Implementation

The algorithms presented in the last two sections can be implemented with two computer programs. The first program would implement the algorithm in Section 5. Its input would be a mask and its output would be a reconstruction algorithm. The second program would implement the program in Section 4. Its input would consist of a reconstruction algorithm and an autocorrelation function and its output would be the object function. With this arrangement, if one wished to reconstruct many object functions using the same mask, the first program would have to be run only once.

Acknowledgement

This research was supported by Air Force Wright Aeronautical Laboratories, Avionics Laboratory, under contract F33615-83-C-1046, DARPA order 5205.

References

1. Yu. M. Bruck and L.G. Sodin, "On the ambiguity of the image reconstruction problem," Opt. Commun. 30, 304-308 (1979).
2. M.A. Fiddy, B.J. Brames, and J.C. Dainty, "Enforcing irreducibility for phase retrieval in two dimensions," Opt. Lett. 8, 96-98 (1983).
3. J.R. Fienup, "Reconstruction of objects having latent reference points," J. Opt. Soc. Am. 73, 1421-1426 (1983).

Reprinted from Journal of the Optical Society of America A, Vol. 4, page 124, January 1987

Phase retrieval for discrete functions with support constraints

Thomas R. Crimmins

Environmental Research Institute of Michigan, P.O. Box 8618, Ann Arbor, Michigan 48107-8618

Received May 6, 1986; accepted August 15, 1986

It is shown that phase retrieval for two-dimensional discrete object functions satisfying certain support constraints is unique among all object functions satisfying the same support constraint. Closed-form reconstruction algorithms for reconstructing such object functions from their autocorrelation functions are defined. Also, an algorithm for generating these reconstruction algorithms is described.

1. INTRODUCTION

The reconstruction of object functions having nonredundant spacings was discussed previously.[1] Hayes and Quatieri[2] showed that the boundaries of triangular objects can be reconstructed by making use of certain spacings in the object that are nonredundant. In another direction, Bruck and Sodin[3] showed that the uniqueness of phase retrieval is equivalent to the irreducibility of a polynomial in two variables that is closely related to the Fourier transform (z transform) of the object function. Fiddy *et al.*[4] used Eisenstein's irreducibility criterion to prove uniqueness for object functions satisfying certain support constraints and showed that Fienup's input-output iterative Fourier-transform algorithm[5-7] converged faster to a better reconstruction when these constraints were satisfied. Fienup[8] presented a closed-form algorithm for reconstructing such object functions from their autocorrelation functions. He also presented a similar closed-form reconstruction algorithm for objects satisfying a triangular support constraint and thereby showed that such objects are uniquely defined by their autocorrelation functions among all object functions satisfying the same support constraint.

A generalization of Fienup's results to a wider class of support constraints is presented here. Also, an algorithm for generating closed-form reconstruction algorithms is described. Brames[9] recently obtained a result similar to the uniqueness theorem described in Section 3.

2. MASKS

Let $\mathcal{R}^2$ denote the Euclidean plane, and let Z^2 denote the points in $\mathcal{R}^2$ with integer coordinates. A finite subset of Z^2 is a *mask* if it contains at least three noncollinear points and its convex hull in $\mathcal{R}^2$ (the smallest convex set containing it) has no parallel sides. These conditions on a mask are needed in the proof of the uniqueness theorem in the next section (proof in Appendix A) and in the proof of the algorithm presented in Section 5 (proof in Appendix C). Let M be a mask and let $[M]$ denote its convex hull in $\mathcal{R}^2$. $[M]$ is then a convex polygon (including its interior; see Fig. 1). A vertex v of $[M]$ is *opposite* a side s of $[M]$ if the line through v and parallel to s contains no points of $[M]$ other than v (see Fig. 2). A vertex of $[M]$ is a *reference point* of M if it is opposite some side of $[M]$ (see Fig. 3). The set of all reference points of M will be denoted by $R(M)$.

3. UNIQUENESS THEOREM

Let f be a complex-valued function on Z^2. The support of a function on Z^2 is the set of points at which the function is nonzero. Let $\mathcal{S}(f)$ denote the support of f. If $\mathcal{S}(f)$ is a finite set, the autocorrelation function of f is defined for $x \in Z^2$ by

$$r(x) = \sum_{y \in Z^2} f(y) f(y - x)^*, \tag{1}$$

where the * denotes complex conjugation. Let f_1 be another complex-valued function on Z^2 with finite support $\mathcal{S}(f_1)$ and autocorrelation function r_1. We have the following uniqueness theorem:

Theorem: If M is a mask, $R(M) \subseteq \mathcal{S}(f) \subseteq M$, $\mathcal{S}(f_1) \subseteq M$, and $r = r_1$, then there exists a complex number α of modulus 1 such that $f_1 = \alpha f$.

The proof is in Appendix A.

4. RECONSTRUCTION ALGORITHMS

In this section closed-form algorithms for reconstructing a function from its autocorrelation function will be described.

Let S be the number of vertices of $[M]$. Let $v_0, \ldots, v_{S-1}$ be an ordering of the vertices going around $[M]$ in the counterclockwise direction, and let $p_0, \ldots, p_{T-1}$ be a similar ordering of the reference points of M. By lemma A2 in Appendix A, $R(M)$ contains an odd number of points so that T is odd. Let $K = (T-1)/2$ and let $q_n = p_{(nK) \bmod T}$ for $n = 0, \ldots, T-1$. Since K and T are relatively prime, the q_n are distinct and hence run through all the points of $R(M)$ (see Fig. 4). By lemma A4 in Appendix A, q_n and $q_{(n+1) \bmod T}$ have unique separation in M. That is, if $x, y \in M$ and $x - y = q_{(n+1) \bmod T} - q_n$, then $x = q_{(n+1) \bmod T}$ and $y = q_n$.

Let N be the number of points in M. A *reconstruction algorithm* for the mask M is an ordered pair (q, m), for which $q = (q_0, \ldots, q_{N-1})$ is an ordering of the points in M and $m = (m_T, \ldots, m_{N-1})$ is a sequence of integers satisfying the fol-

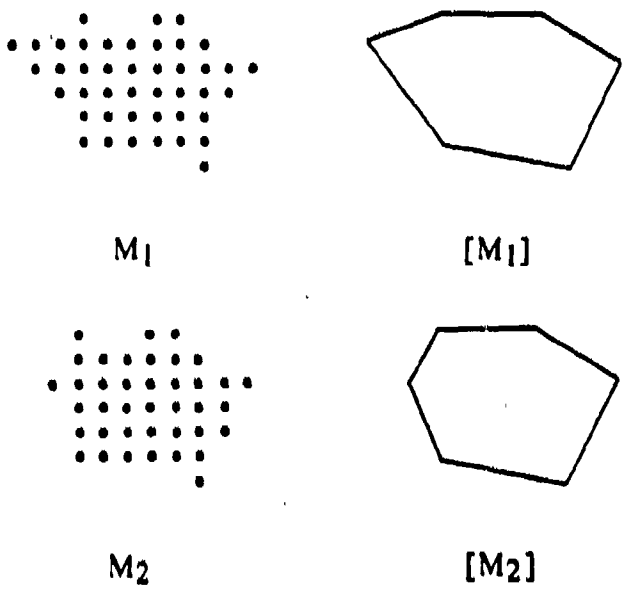


Fig. 1. M_1 is a mask. M_2 is not a mask, since $[M_2]$ has two parallel sides.

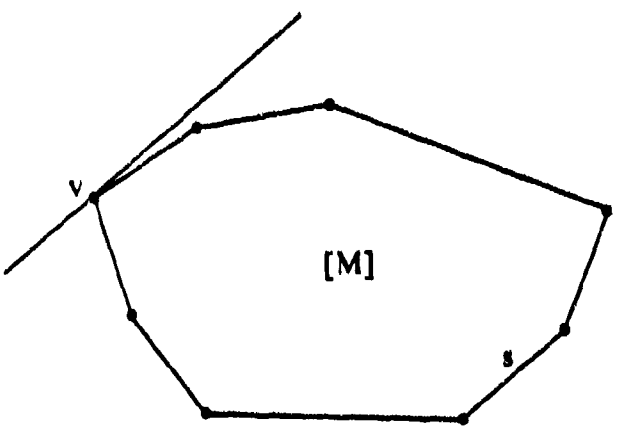


Fig. 2. The vertex v is opposite the side s.

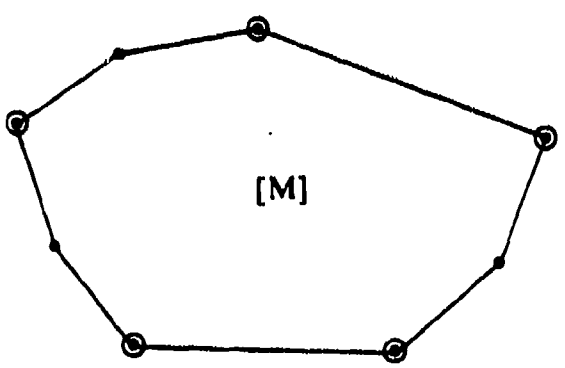


Fig. 3. The circled vertices of $[M]$ are the reference points of the mask M.

lowing conditions. The points $q_0, \ldots, q_{T-1}$ are as described above. For $n = T, \ldots, N-1$, the integers m_n satisfy the conditions $0 \le m_n \le T-1$, and $M \cap (M + q_n - q_{m_n}) \subseteq [q_0, \ldots, q_n]$ and $M \cap (M - q_n + q_{m_n}) \subseteq [q_0, \ldots, q_{n-1}]$. In the next section, an algorithm for generating such reconstruction algorithms will be described.

In order to justify the above definition of reconstruction algorithms, it will now be shown how they can be used to reconstruct a function from its autocorrelation function.

Let f be a complex-valued function on Z^2 satisfying $R(M) \subseteq \mathcal{S}(f) \subseteq M$, and let r be its autocorrelation function. Let $x = q_{(n+1) \bmod T} - q_n$, and suppose that for some $y \in Z^2$, $f(y)f(y-x)^* \neq 0$; then $y \in \mathcal{S}(f) \subseteq M$ and $y - x \in \mathcal{S}(f) \subseteq M$. Also, $y - (y-x) = x = q_{(n+1) \bmod T} - q_n$. Since $q_{(n+1) \bmod T}$ and q_n have unique separation in M, it follows that $y = q_{(n+1) \bmod T}$ and $y - x = q_n$. Therefore $y = q_{(n+1) \bmod T}$ is the only $y \in Z^2$ for which $f(y)f(y-x)^* \neq 0$; hence

$$r(q_{(n+1) \bmod T} - q_n) = f(q_{(n+1) \bmod T}) f(q_n)^*, \tag{2}$$

and since $R(M) \subseteq \mathcal{S}(f)$, $r(q_{(n+1) \bmod T} - q_n) \neq 0$. It now follows from Eq. (2) that

$$|f(q_0)|^2 = \frac{\prod_{n=0}^{K} r(q_{2n} - q_{(2n+1) \bmod T})}{\prod_{n=0}^{K-1} r(q_{2n+2} - q_{2n+1})}. \tag{3}$$

Since f is defined by r only up to multiplication by a modulus 1 complex number, we may require that $f(q_0) > 0$. $f(q_0)$ is then equal to the positive square root of the right-hand side of Eq. (3). $f(q_n)$ can then be computed for $n = 1, \ldots, T-1$ from the formula $f(q_n) = r(q_n - q_{n-1})/f(q_{n-1})^*$. It is shown in Appendix B that if (q, m) is a reconstruction algorithm, then the following program will compute $f(q_n)$ for $n = T, \ldots, N-1$. Set $f(x) = 0$ for $x \in Z^2$ and $x \neq q_n$, $n = 0, \ldots, T-1$, and set $n = T-1$.

Step 1. If $n = N-1$, stop. Otherwise, $n \leftarrow n+1$.

Step 2. $f(q_n)$

$$= \left[r(q_n - q_{m_n}) - \sum_{k=0}^{n-1} f(q_k) f(q_k - q_n + q_{m_n})^* \right] \Big/ f(q_{m_n})^*.$$

Step 3. Go to step 1.

5. ALGORITHM FOR GENERATING RECONSTRUCTION ALGORITHMS

It will be assumed that we are given a sequence of all vertices $v_0, \ldots, v_{S-1}$ of M, where M is a mask and the sequence is ordered in the counterclockwise direction around $[M]$.

For $n = 0, \ldots, S-1$, let s_n be the side of $[M]$ with endpoints v_n and $v_{(n+1) \bmod S}$. Let U be the linear operator on $\mathcal{R}^2$ that rotates each vector in $\mathcal{R}^2$ 90° counterclockwise.

First, the reference points $p_0, \ldots, p_{T-1}$ must be found.

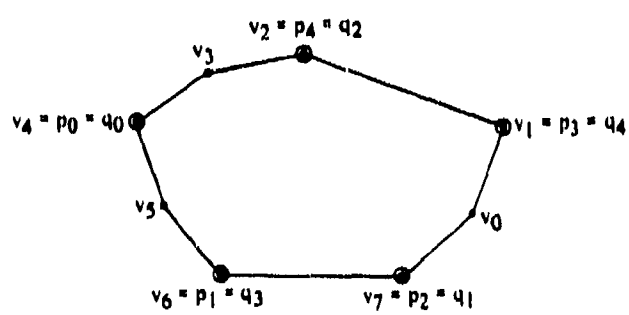


Fig. 4. The numberings of the vertices and reference points of a mask. Here $S = 8$, $T = 5$, and $K = 2$.

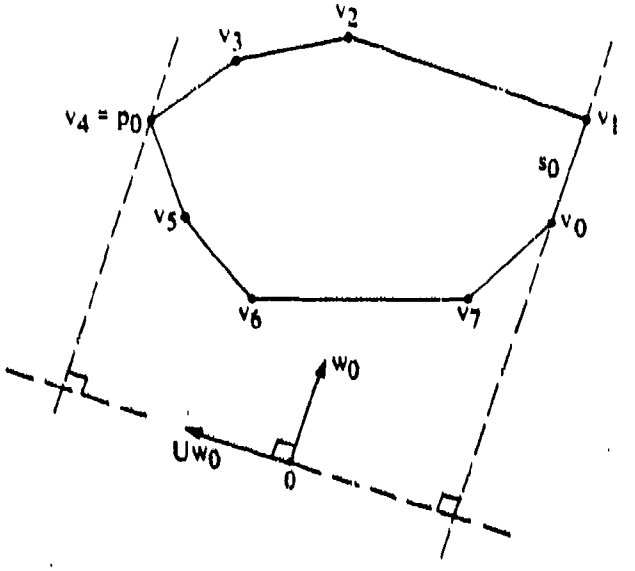


Fig. 5. An illustration of the vectors w_n and Uw_n. Here $n = 0$ and the origin in $\mathcal{R}^2$ is denoted by 0.

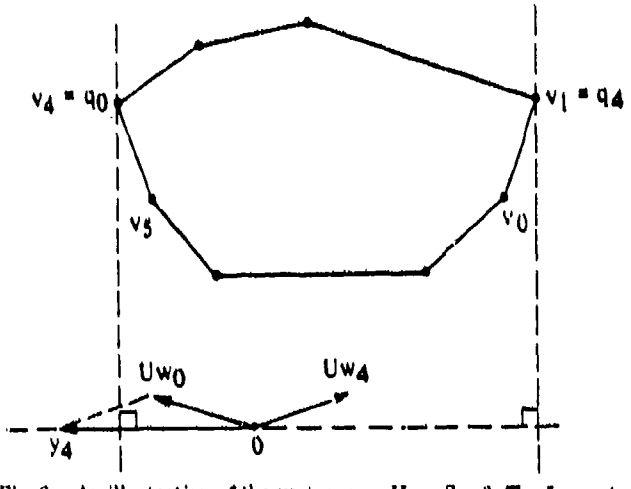


Fig. 6. An illustration of the vectors y_n. Here $S = 8$, $T = 5$, $n = 4$, $k_4 = 1$, $(k_4 - 1) \bmod S = 0$, $(n + 1) \bmod T = 0$, and $k_0 = 4$.

Note that every side of $[M]$ has a vertex opposite it that is therefore a reference point. Of course, several sides may have the same vertex opposite them. Let $w_n = v_{(n+1) \bmod S} - v_n$ for $n = 0, \ldots, S - 1$. A vertex v_m is opposite a side s_n if and only if $\langle v_m, Uw_n \rangle \geq \langle v_k, Uw_n \rangle$ for $k = 0, \ldots, S - 1$, where $\langle\ \rangle$ denotes the inner product on $\mathcal{R}^2$ (see Fig. 5). Thus, by taking each side in the order $s_0, \ldots, s_{S-1}$, all the reference points of M will be found, and if they are numbered in the order in which they are found, $p_0, \ldots, p_{T-1}$, then the ordering will be in the counterclockwise direction around $[M]$.

As mentioned above, T is odd. Let $K = (T - 1)/2$ and $q_n = p_{(nK) \bmod T}$, $n = 0, \ldots, T - 1$. Since each q_n is a reference point and therefore is a vertex of $[M]$, there exists an integer k_n such that $0 \leq k_n \leq S - 1$ and $q_n = v_{k_n}$. For $n = 0, \ldots, T - 1$, define

$$y_n = Uw_{(k_n - 1) \bmod S} - Uw_{k_{(n+1) \bmod T}}; \tag{4}$$

then by lemma A3 in Appendix A, for $x \in M$, $x \neq q_n$, and $x \neq q_{(n+1) \bmod T}$, $\langle q_n, y_n \rangle < \langle x, y_n \rangle < \langle q_{(n+1) \bmod T}, y_n \rangle$. This is equivalent to saying that all points in M excluding q_n and $q_{(n+1) \bmod T}$ lie strictly between lines perpendicular to y_n and passing through q_n and $q_{(n+1) \bmod T}$ (see Fig. 6). (The uniqueness of separation of q_n and $q_{(n+1) \bmod T}$ mentioned in Section 4 follows from this double inequality.)

Let $\alpha_n = q_n + q_{(n+1) \bmod T}$, and let $\beta_n = \alpha_n/2$ for $n = 0, \ldots, T - 1$. β_n is then the midpoint of the line segment joining q_n and $q_{(n+1) \bmod T}$. Let $D = M \setminus R(M)$ (set difference), and let ϕ be the characteristic function of D as a subset of Z^2; i.e., ϕ is the function on Z^2 which is 1 on D and 0 outside D. For $x \in D$ and $0 \leq n \leq T - 1$, let $h_n(x) = \langle x - \beta_n, y_n \rangle$; then $|h_n(x)|/\|y_n\|$ is the distance from x to the line through β_n and perpendicular to y_n, where $\|y_n\|$ denotes the Euclidean norm of y_n (see Fig. 7). The set D contains $N - T$ points, and we define T orderings of the points in D. $D = \{d_{n,0}, \ldots, d_{n,N-T-1}\}$, $n = 0, \ldots, T - 1$, satisfying $|h_n(d_{n,k})| \geq |h_n(d_{n,k+1})|$ for $k = 0, \ldots, N - T - 2$. The following program generates sequences $q_T, \ldots, q_{N-1}$ and $m_T, \ldots, m_{N-1}$. Set $n = T - 1$ and $k = 0$ and enter the following loop:

Step 1. If $n = N - 1$, stop. Otherwise, define $b = \min\{j: 0 \leq j \leq N - T - 1 \text{ and } \phi(d_{k,j}) = 1\}$.
Step 2. If $\phi(\alpha_k - d_{k,b}) = 1$, go to step 7.
Step 3. $n \leftarrow n + 1$.
Step 4. Define $q_n = d_{k,b}$.
Step 5. If $h_k(q_n) \geq 0$, define $m_n = k$. Otherwise, define $m_n = (k + 1) \bmod T$.
Step 6. $\phi(q_n) \leftarrow 0$.
Step 7. $k \leftarrow (k + 1) \bmod T$ and go to step 1.

It is shown in Appendix C that the loop is not infinite, and if $q = (q_0, \ldots, q_{N-1})$ and $m = (m_T, \ldots, m_{N-1})$ then (q, m) is a reconstruction algorithm.

An example of a reconstruction algorithm generated by the algorithm described above is illustrated in Figs. 8 and 9. Figure 8 shows a mask, a numbering of its vertices, and the resulting numberings of its reference points. Here $w_0 =$

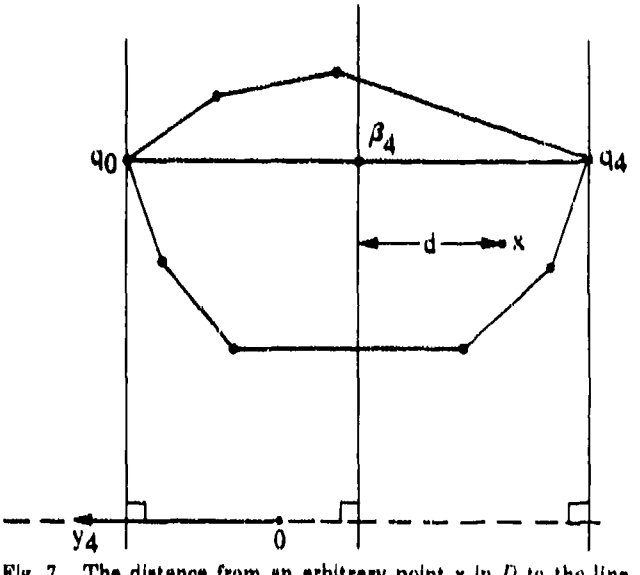


Fig. 7. The distance from an arbitrary point x in D to the line through β and perpendicular to y_n is $d = |h_n(x)|/\|y_n\|$. Here $S = 8$, $T = 5$, and $n = 4$.

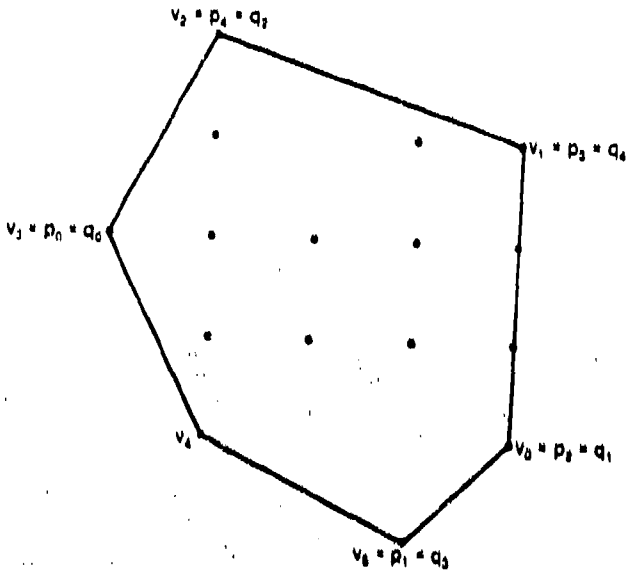


Fig. 8. Example of a mask, a numbering of its vertices, and the resulting numberings of its reference points.

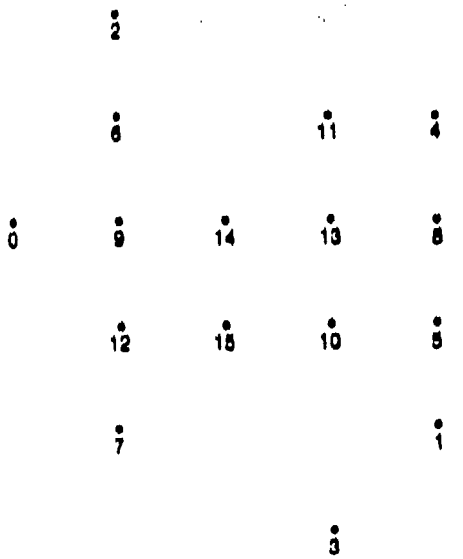


Fig. 9. A reconstruction algorithm (q, m) generated for the mask and vertex numbering shown in Fig. 8. The numbers under the points of the mask indicate the ordering $q = (q_0, \ldots, q_{15})$ and $m = (0, 1, 4, 0, 1, 2, 3, 4, 0, 1, 2)$.

$(0, 3)$, $w_1 = (-3, 1)$, $w_2 = (-1, -2)$, $w_3 = (1, -2)$, $w_4 = (2, -1)$, and $w_5 = (1, 1)$. Also, $y_0 = U(w_2 - w_0) = (5, -1)$, $y_1 = U(w_5 - w_3) = (-3, 2)$, $y_2 = U(w_1 - w_5) = (0, -4)$, $y_3 = U(w_4 - w_1) = (2, 5)$, and $y_4 = U(w_0 - w_3) = (-5, -1)$. The reconstruction algorithm (q, m) that is generated is illustrated in Fig. 9.

6. IMPLEMENTATION

The algorithms presented in the last two sections can be implemented with two computer programs. The first program would implement the algorithm in Section 5. Its input would be a mask, and its output would be a reconstruction algorithm. The second program would implement the program in Section 4. Its input would consist of a reconstruction algorithm and an autocorrelation function, and its output would be the object function. With this arrangement, if one wished to reconstruct many object functions using the same mask, the first program would have to be run only once.

7. CONCLUSIONS

It has been shown that if a function is zero outside a given mask and is nonzero at the reference points of the mask, then it is uniquely determined (up to multiplication by a complex number with modulus 1) by its autocorrelation function among all other object functions that are zero outside the mask. (A mask is a set of points in the discrete lattice whose convex hull has no parallel sides.) Moreover, there is an algorithm for generating reconstruction algorithms for any given mask, which in turn can be used to reconstruct object functions satisfying the above-mentioned conditions from their autocorrelation functions.

This theory has some similarity to holography.[10,11] However, several (at least three) reference points are used here, whereas only one reference point is needed in the holographic situation. On the other hand, the holographic reference point must be isolated from the rest of the object, whereas no such isolation of the reference points is required here. It is interesting to speculate whether there might be a more general theory of which this theory and holography would both be special cases.

APPENDIX A

Let S, T, v_n for $n = 0, \ldots, S-1$ and q_n for $n = 0, \ldots, T-1$ be defined as in Section 4. Therefore all the q_n's referred to in this appendix are reference points. Also, let U, s_n, and w_n for $n = 0, \ldots, S-1$ and k_n and y_n for $n = 0, \ldots, T-1$ be defined as in Section 5. Let t_n be the side of $[R(M)]$ with endpoints p_n and $p_{(n+1) \bmod T}$ (see Fig. 10), and let $u_n = p_{(n+1) \bmod T} - p_n$ for $n = 0, \ldots, T-1$. We note for future reference that for $v, w \in \mathcal{R}^2$, $\langle v, Uv \rangle = 0$, $U^2 v = -v$, $\langle Uv, Uw \rangle = \langle v, w \rangle$, and $\langle v, Uw \rangle = \langle Uv, U^2 w \rangle = -\langle Uv, w \rangle$. The proof of the uniqueness theorem in Section 3 requires a series of lemmas.

Lemma A1

$R(M)$ is a mask and $R(R(M)) = R(M)$.

Proof

Suppose it can be shown that every vertex of $[R(M)]$ is opposite some side of $[R(M)]$. Since at most one vertex can be opposite a given side and the number of vertices equals

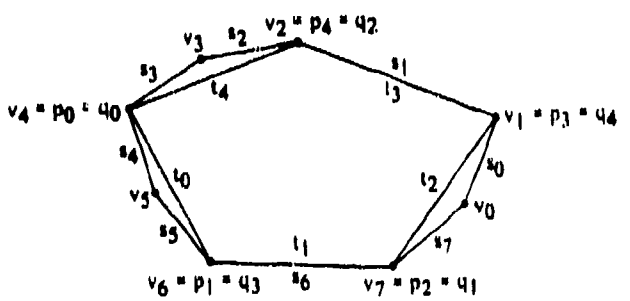


Fig. 10. The set $[R(M)]$ is the convex polygon with sides t_i, $i = 0, \ldots, 4$.

the number of sides, it would then follow that every side must have a vertex opposite it, and therefore no two sides can be parallel; hence $R(M)$ is a mask. Also, since every vertex is opposite a side, $R(R(M))$ is the set of all vertices of $[R(M)]$, which is equal to the set $R(M)$; hence $R(R(M)) = R(M)$. Thus it suffices to show that every vertex of $[R(M)]$ is opposite some side of $[R(M)]$.

The vertices of $[R(M)]$ are p_m, $m = 0, \ldots, T-1$. Let m be fixed but arbitrary, $0 \le m \le T-1$; then p_m is also a vertex of $[M]$, and hence $p_m = v_k$ for some k. Let v_a be the vertex of $[M]$ opposite side $s_{(k-1) \bmod S}$ of $[M]$, and let v_b be the vertex of $[M]$ opposite side s_k of $[M]$; then v_a and v_b are in $R(M)$, and $v_a = p_n$ for some n. (Refer to Fig. 10 and take $m = 0$. In this case $k = 4$, $a = 7$, $b = 1$ and $n = 2$.) If v_a and v_b are the same vertex, then there can be no side of $[M]$ opposite v_k, but $v_k = p_m \in R(M)$, and so v_k must be opposite some side of $[M]$. Therefore $v_a \ne v_b$. It then follows that $v_b = p_{(n+1) \bmod T}$. It will be shown that p_m is opposite side t_n of $[R(M)]$. That is, we wish to show that for $0 \le j \le T-1$ and $j \ne m$,

$$\langle p_j, Uu_n \rangle < \langle p_m, Uu_n \rangle. \tag{A1}$$

Since v_a is opposite side $s_{(k-1) \bmod S}$ of $[M]$, it follows that for $0 \le i \le S-1$ and $i \ne a$,

$$\langle Uw_{(k-1) \bmod S}, v_i - v_a \rangle < 0, \tag{A2}$$

and since v_b is opposite s_k, for $0 \le i \le S-1$ and $i \ne b$,

$$\langle Uw_k, v_i - v_b \rangle < 0. \tag{A3}$$

By setting $i = b$ in expression (A2), we obtain $\langle Uw_{(k-1) \bmod S}, v_b - v_a \rangle < 0$, and thus

$$\begin{aligned} \langle v_{(k-1) \bmod S} - p_m, Uu_n \rangle &= \langle v_{(k-1) \bmod S} - v_k, U(p_{(n+1) \bmod T} - p_n) \rangle \\ &= -\langle w_{(k-1) \bmod S}, U(v_b - v_a) \rangle \\ &= \langle Uw_{(k-1) \bmod S}, v_b - v_a \rangle \\ &< 0. \end{aligned} \tag{A4}$$

By setting $i = a$ in expression (A3), we obtain $\langle Uw_k, v_a - v_b \rangle < 0$, and thus

$$\begin{aligned} \langle v_{(k+1) \bmod S} - p_m, Uu_n \rangle &= \langle v_{(k+1) \bmod S} - v_k, U(p_{(n+1) \bmod T} - p_n) \rangle \\ &= \langle w_k, U(v_b - v_a) \rangle \\ &= \langle Uw_k, v_a - v_b \rangle \\ &< 0. \end{aligned} \tag{A5}$$

Since $v_{(k-1) \bmod S}$, $v_k (= p_m)$ and $v_{(k+1) \bmod S}$ are distinct vertices of $[M]$, the vectors $v_{(k-1) \bmod S} - p_m$ and $v_{(k+1) \bmod S} - p_m$ are linearly independent. Let $p \in R(M)$, $p \ne v_{(k-1) \bmod S}, p_m, v_{(k+1) \bmod S}$. There then exist real numbers α and β, such that

$$p - p_m = \alpha(v_{(k-1) \bmod S} - p_m) + \beta(v_{(k+1) \bmod S} - p_m). \tag{A6}$$

Also, since $\langle v_k, Uw_k \rangle < \langle p, Uw_k \rangle$,

$$\begin{aligned} 0 &< \langle p - v_k, Uw_k \rangle \\ &= \langle p - p_m, U(v_{(k+1) \bmod S} - p_m) \rangle \\ &= \alpha\langle v_{(k-1) \bmod S} - p_m, U(v_{(k+1) \bmod S} - p_m) \rangle \\ &\quad + \beta\langle v_{(k+1) \bmod S} - p_m, U(v_{(k+1) \bmod S} - p_m) \rangle \\ &= \alpha\langle v_{(k-1) \bmod S} - p_m, U(p_m - v_{(k-1) \bmod S}) \rangle \\ &= -\alpha\langle w_{(k-1) \bmod S}, U(v_{(k+1) \bmod S} - v_k) \rangle \\ &= \alpha\langle Uw_{(k-1) \bmod S}, v_{(k+1) \bmod S} - v_k \rangle. \end{aligned} \tag{A7}$$

Since $\langle Uw_{(k-1) \bmod S}, v_{(k+1) \bmod S} - v_k \rangle > 0$, it follows from expression (A7) that $\alpha > 0$. Similarly,

$$\begin{aligned} 0 &< \langle p - v_k, Uw_{(k-1) \bmod S} \rangle \\ &= \langle p - p_m, U(p_m - v_{(k-1) \bmod S}) \rangle \\ &= \alpha\langle v_{(k-1) \bmod S} - p_m, U(p_m - v_{(k-1) \bmod S}) \rangle \\ &\quad + \beta\langle v_{(k+1) \bmod S} - p_m, U(p_m - v_{(k-1) \bmod S}) \rangle \\ &= -\alpha\langle p_m - v_{(k-1) \bmod S}, U(p_m - v_{(k-1) \bmod S}) \rangle \\ &\quad + \beta\langle v_{(k+1) \bmod S} - p_m, U(p_m - v_{(k-1) \bmod S}) \rangle \\ &= \beta\langle v_{(k+1) \bmod S} - p_m, U(p_m - v_{(k-1) \bmod S}) \rangle \\ &= \beta\langle v_{(k+1) \bmod S} - v_k, Uw_{(k-1) \bmod S} \rangle. \end{aligned} \tag{A8}$$

Since $\langle v_{(k+1) \bmod S} - v_k, Uw_{(k-1) \bmod S} \rangle > 0$, it follows from expression (A8) that $\beta > 0$. Using expressions (A4), (A5), and (A6) and the fact that $\alpha > 0$ and $\beta > 0$, we have, for $p \ne v_{(k-1) \bmod S}, p_m, v_{(k+1) \bmod S}$,

$$\begin{aligned} \langle p - p_m, Uu_n \rangle &= \alpha\langle v_{(k-1) \bmod S} - p_m, Uu_n \rangle \\ &\quad + \beta\langle v_{(k+1) \bmod S} - p_m, Uu_n \rangle \\ &< 0. \end{aligned} \tag{A9}$$

Inequality (A1) now follows from expressions (A4), (A5), and (A9). This completes the proof of lemma A1.

Lemma A2

The number T of points in $R(M)$ is odd, and if $K = (T-1)/2$ and $0 \le j \le T-1$ then p_j, as a vertex of $[R(M)]$, is opposite side $t_{(K+j) \bmod T}$ of $[R(M)]$.

Proof

It follows from lemma A1 that every side of $[R(M)]$ has exactly one vertex of $[R(M)]$ opposite it and every vertex; i.e., every point in $R(M)$, is opposite exactly one side. Thus there is a positive integer $K \le T-2$, such that p_{K+1} is opposite side t_0. p_1 is then opposite side t_{K+1}, and p_2 is opposite side $t_{(K+2) \bmod T}$. More generally, p_j is opposite side $t_{(K+j) \bmod T}$. Setting $j = T-K$, we find that p_{T-K} is opposite t_0. However, p_{K+1} is opposite t_0. Therefore $T-K = K+1$ or $T = 2K+1$. This completes the proof of lemma A2.

Lemma A3

For $x \in [M]$, $x \ne q_j, q_{(j+1) \bmod T}$, $j = 0, \ldots, T-1$, $\langle q_j, y_j \rangle < \langle x, y_j \rangle < \langle q_{(j+1) \bmod T}, y_j \rangle$.

Proof

It suffices to show that the inequalities hold for all vertices v of $[M]$, $v \ne q_j, q_{(j+1) \bmod T}$. Let j be fixed but arbitrary. For convenience let $m = (k_j - 1) \bmod S$ and $n = k_{(j+1) \bmod T}$; then $y_j = Uw_m - Uw_n$.

First we will show that $q_{(j+1) \bmod T}$, as a vertex of $[M]$, is opposite side s_m of $[M]$. Let v_a be the vertex of $[M]$ opposite side s_m of $[M]$, and let v_b be opposite $s_{(m+1) \bmod S}$; then v_a and

v_h are in $R(M)$, and $v_n = p_k$ for some k. [Refer to Fig. 10 and take $j = 4$; then $k_j = 1$, $m = 0$, $(j+1) \bmod T = 0$, $n = k_0 = 4$, $a = 4$, $b = 0$, and $k = 0$.] By the argument in the proof of lemma A1, $v_h = p_{(k+1) \bmod r}$, and $p_{(jK) \bmod r} (= q_j = v_{(m+1) \bmod S})$, as a vertex of $[R(M)]$, is opposite side t_k of $[R(M)]$. By lemma A2, $p_{(jK) \bmod r}$ is opposite $t_{(j+1)K \bmod r}$. Hence, by lemma A1, $t_k = t_{(j+1)K \bmod r}$ and $k = (j+1)K \bmod T$. Thus $q_{(j+1) \bmod T} = p_{(j+1)K \bmod r} = p_k = v_n$, and therefore $q_{(j+1) \bmod T}$, as a vertex of $[M]$, is opposite side s_m of $[M]$.

By a similar argument it can be shown that q_j, as a vertex of $[M]$, is opposite side s_n of $[M]$.

Since $q_{(j+1) \bmod T}$ is opposite s_m and $v = q_{(j+1) \bmod T}$, it follows that $\langle q_{(j+1) \bmod T}, Uw_m \rangle \geq \langle v, Uw_m \rangle$ or $\langle q_{(j+1) \bmod T} - v, Uw_m \rangle > 0$. Also, since $v \in [M]$, $\langle v - v_n, Uw_n \rangle \geq 0$. Therefore

$$\begin{aligned}
&\langle q_{(j+1) \bmod T}, y_j \rangle - \langle v, y_j \rangle \\
&\quad = \langle q_{(j+1) \bmod T} - v, y_j \rangle \\
&\quad = \langle q_{(j+1) \bmod T} - v, Uw_m - Uw_n \rangle \\
&\quad = \langle q_{(j+1) \bmod T} - v, Uw_m \rangle - \langle q_{(j+1) \bmod T} - v, Uw_n \rangle \\
&\quad = \langle q_{(j+1) \bmod T} - v, Uw_m \rangle - \langle v_n - v, Uw_n \rangle \\
&\quad = \langle q_{(j+1) \bmod T} - v, Uw_m \rangle + \langle v - v_n, Uw_n \rangle \\
&\quad > 0. \qquad \text{(A10)}
\end{aligned}$$

Since q_j is opposite s_n and $v = q_j$, $\langle q_j, Uw_n \rangle \geq \langle v, Uw_n \rangle$ or $\langle q_j - v, Uw_n \rangle > 0$. Also, since $v \in [M]$, $\langle v - v_{(m+1) \bmod S}, Uw_n \rangle \geq 0$. Therefore

$$\begin{aligned}
\langle v, y_j \rangle - \langle q_j, y_j \rangle &= \langle v - q_j, y_j \rangle \\
&= \langle v - q_j, Uw_m - Uw_n \rangle \\
&= \langle v - q_j, Uw_m \rangle + \langle q_j - v, Uw_n \rangle \\
&= \langle v - v_{k_j}, Uw_m \rangle + \langle q_j - v, Uw_n \rangle \\
&= \langle v - v_{(n+1) \bmod S}, Uw_m \rangle + \langle q_j - v, Uw_n \rangle \\
&> 0. \qquad \text{(A11)}
\end{aligned}$$

It now follows from expressions (A10) and (A11) that $\langle q_j, y_j \rangle < \langle v, y_j \rangle < \langle q_{(j+1) \bmod T}, y_j \rangle$. This completes the proof of lemma A3.

In the remainder of this appendix, all modulo arithmetic will be mod T. For convenience, we define $m \oplus n = (m+n) \bmod T$.

The next lemma asserts that q_j and $q_{j \oplus 1}$ have unique separation in M.

Lemma A4
For $0 \leq j \leq T-1$, if $x_1, x_2 \in M$ and $x_1 - x_2 = q_{j \oplus 1} - q_j$, then $x_1 = q_{j \oplus 1}$ and $x_2 = q_j$.

Proof
If either $x_1 \neq q_{j \oplus 1}$ or $x_2 \neq q_j$, then it follows from lemma A3 that

$$\begin{aligned}
\langle x_1 - x_2, y_j \rangle &= \langle x_1, y_j \rangle - \langle x_2, y_j \rangle \\
&< \langle q_{j \oplus 1}, y_j \rangle - \langle q_j, y_j \rangle \\
&= \langle q_{j \oplus 1} - q_j, y_j \rangle. \qquad \text{(A12)}
\end{aligned}$$

which contradicts the assumption that $x_1 - x_2 = q_{j \oplus 1} - q_j$. Therefore $x_1 = q_{j \oplus 1}$ and $x_2 = q_j$. This completes the proof of lemma A4.

Let g and h be complex-valued functions on Z^2, and let $g * h$ denote the convolution of g and h; that is, if $f = g * h$, then

$$f(x) = \sum_{u \in Z^2} g(u) h(x-u). \qquad \text{(A13)}$$

We define $\mathcal{S}(g) + \mathcal{S}(h) = \{x + y: x \in \mathcal{S}(g) \text{ and } y \in \mathcal{S}(h)\}$. The following lemma is fundamental.

Lemma A5
If $f = g * h$, then $[\mathcal{S}(f)] = [\mathcal{S}(g) + \mathcal{S}(h)]$.

Proof
It follows from Eq. (A13) that $\mathcal{S}(f) \subseteq \mathcal{S}(g) + \mathcal{S}(h)$; hence $[\mathcal{S}(f)] \subseteq [\mathcal{S}(g) + \mathcal{S}(h)]$. It remains to be shown that $[\mathcal{S}(g) + \mathcal{S}(h)] \subseteq [\mathcal{S}(f)]$.

Let x be a vertex of $[\mathcal{S}(g) + \mathcal{S}(h)]$; then there exists a $y \in \mathcal{R}^2$ such that for $x' \in \mathcal{S}(g) + \mathcal{S}(h)$ and $x' \neq x$,

$$\langle x', y \rangle < \langle x, y \rangle. \qquad \text{(A14)}$$

Also, since x is a vertex of $[\mathcal{S}(g) + \mathcal{S}(h)]$, it follows that $x \in \mathcal{S}(g) + \mathcal{S}(h)$, and hence there exist $x_1 \in \mathcal{S}(g)$ and $x_2 \in \mathcal{S}(h)$ such that $x = x_1 + x_2$. We will show that this decomposition of x is unique. Suppose that $x = x_1' + x_2'$ with $x_1' \in \mathcal{S}(g)$ and $x_2' \in \mathcal{S}(h)$; then

$$\begin{aligned}
\langle x_1, y \rangle + \langle x_2, y \rangle &= \langle x_1 + x_2, y \rangle \\
&= \langle x, y \rangle \\
&= \langle x_1' + x_2', y \rangle \\
&= \langle x_1', y \rangle + \langle x_2', y \rangle. \qquad \text{(A15)}
\end{aligned}$$

Therefore either $\langle x_1', y \rangle \geq \langle x_1, y \rangle$ or $\langle x_2', y \rangle \geq \langle x_2, y \rangle$ or both. Suppose that $\langle x_1', y \rangle \geq \langle x_1, y \rangle$. Let $x' = x_1' + x_2$; then $x' \in \mathcal{S}(g) + \mathcal{S}(h)$ and

$$\begin{aligned}
\langle x', y \rangle &= \langle x_1', y \rangle + \langle x_2, y \rangle \\
&\geq \langle x_1, y \rangle + \langle x_2, y \rangle \\
&= \langle x, y \rangle. \qquad \text{(A16)}
\end{aligned}$$

Therefore, by inequality (A14), $x' = x$, which implies that $x_1' = x_1$ and hence $x_2' = x_2$. If $\langle x_2', y \rangle \geq \langle x_2, y \rangle$, a similar argument leads to the same conclusion. Therefore the decomposition $x = x_1 + x_2$ with $x_1 \in \mathcal{S}(g)$ and $x_2 \in \mathcal{S}(h)$ is unique. Suppose, for a particular $u_0 \in Z^2$, that $g(u_0)h(x-u_0) \neq 0$; then $u_0 \in \mathcal{S}(g)$, $x - u_0 \in \mathcal{S}(h)$ and $x = u_0 + (x - u_0)$. By the uniqueness of the decomposition of x it follows that $u_0 = x_1$ and $x - u_0 = x_2$. Therefore $f(x) = g(x_1)h(x_2) \neq 0$ and $x \in \mathcal{S}(f)$. Since x was an arbitrary vertex of $[\mathcal{S}(g) + \mathcal{S}(h)]$, it follows that all the vertices of $[\mathcal{S}(g) + \mathcal{S}(h)]$ are in $\mathcal{S}(f)$, and therefore $[\mathcal{S}(g) + \mathcal{S}(h)] \subseteq [\mathcal{S}(f)]$. This completes the proof of lemma A5.

We are now ready to prove the theorem.

Proof of Theorem
Since $r = r_1$, it follows from the results of Bruck and Sodin[1] that there exist functions g and h with finite supports and a

vector $d \in Z^2$ such that $f = g * h$ and $f_1(x) = g * h_1(x - d)$, where $h_1(x) = h(-x)^*$ for $x \in Z^2$.

We have $R(M) = \mathcal{S}(f) = \mathcal{S}(g) + \mathcal{S}(h)$. Therefore there exist $a_0, \ldots, a_{T-1} \in \mathcal{S}(g)$ and $b_0, \ldots, b_{T-1} \in \mathcal{S}(h)$ such that

$$q_j = a_j + b_j, \qquad j = 0, \ldots, T-1. \tag{A17}$$

Now let j be fixed but arbitrary, and let $x \in \mathcal{S}(g)$, $x \neq a_j$. We will show that $\langle a_j, y_j \rangle < \langle x, y_j \rangle$. Suppose, to the contrary, that $\langle x, y_j \rangle \leq \langle a_j, y_j \rangle$. Let $x' = x + b_j$; then, by using lemma A5, $x' \in \mathcal{S}(g) + \mathcal{S}(h) \subseteq [\mathcal{S}(g) + \mathcal{S}(h)] = [\mathcal{S}(f)] \subseteq [M]$. Also,

$$\begin{aligned} \langle x', y_j \rangle &= \langle x, y_j \rangle + \langle b_j, y_j \rangle \\ &\leq \langle a_j, y_j \rangle + \langle b_j, y_j \rangle \\ &= \langle q_j, y_j \rangle. \end{aligned} \tag{A18}$$

It now follows from lemma A3 that $x' = q_j$, which implies that $x = a_j$, contradicting the assumption that $x \neq a_j$. Therefore $\langle a_j, y_j \rangle < \langle x, y_j \rangle$. By a similar argument it can be shown that if $x \in \mathcal{S}(g)$ and $x \neq a_{j \oplus 1}$, then $\langle x, y_j \rangle < \langle a_{j \oplus 1}, y_j \rangle$. Thus, if $x \in \mathcal{S}(g)$ and $x \neq a_j, a_{j \oplus 1}$, then

$$\langle a_j, y_j \rangle < \langle x, y_j \rangle < \langle a_{j \oplus 1}, y_j \rangle. \tag{A19}$$

Also, by similar arguments, it can be shown that if $x \in \mathcal{S}(h)$ and $x \neq b_j, b_{j \oplus 1}$, then

$$\langle b_j, y_j \rangle < \langle x, y_j \rangle < \langle b_{j \oplus 1}, y_j \rangle. \tag{A20}$$

Let $\mathcal{S}(g) - \mathcal{S}(h) = \{x - y : x \in \mathcal{S}(g) \text{ and } y \in \mathcal{S}(h)\}$. Since $\mathcal{S}(h_1) = -\mathcal{S}(h)$, $\mathcal{S}(g) + \mathcal{S}(h_1) = \mathcal{S}(g) - \mathcal{S}(h)$. Also, since $f_1(x) = g * h_1(x - d)$, it follows from lemma A5 that

$$[\mathcal{S}(f_1)] = [\mathcal{S}(g) - \mathcal{S}(h)] + d. \tag{A21}$$

We will show that

$$a_j - b_{j \oplus 1} + d \in M. \tag{A22}$$

We have $a_j - b_{j \oplus 1} \in \mathcal{S}(g) - \mathcal{S}(h)$. Let j be fixed but arbitrary, and let $x \in \mathcal{S}(g) - \mathcal{S}(h)$, $x \neq a_j - b_{j \oplus 1}$. Then there exist $x_1 \in \mathcal{S}(g)$ and $x_2 \in \mathcal{S}(h)$ such that $x = x_1 - x_2$. Since $x \neq a_j - b_{j \oplus 1}$, either $x_1 \neq a_j$ or $x_2 \neq b_{j \oplus 1}$ or both. In any case it follows from expressions (A19) and (A20) that

$$\begin{aligned} \langle x, y_j \rangle &= \langle x_1, y_j \rangle - \langle x_2, y_j \rangle \\ &> \langle a_j, y_j \rangle - \langle b_{j \oplus 1}, y_j \rangle \\ &= \langle a_j - b_{j \oplus 1}, y_j \rangle. \end{aligned} \tag{A23}$$

Therefore $a_j - b_{j \oplus 1}$ is a vertex of $[\mathcal{S}(g) - \mathcal{S}(h)]$, and by Eq. (A21), $a_j - b_{j \oplus 1} + d$ is a vertex of $[\mathcal{S}(f_1)]$. Therefore $a_j - b_{j \oplus 1} + d \in \mathcal{S}(f_1) \subseteq M$, and relation (A22) follows.

By a similar argument it can be shown that

$$a_{j \oplus 1} - b_j + d \in M. \tag{A24}$$

Now

$$\begin{aligned} (a_j - b_{j \oplus 1} + d) - (a_{j \oplus 1} - b_j + d) &= (a_j + b_j) - (a_{j \oplus 1} + b_{j \oplus 1}) \\ &= q_j - q_{j \oplus 1}. \end{aligned} \tag{A25}$$

By lemma A4, $a_j - b_{j \oplus 1} + d = q_j = a_j + b_j$, from which we obtain

$$b_j + b_{j \oplus 1} = d. \tag{A26}$$

From $b_j + b_{j \oplus 1} = d$ and $b_{j \oplus 1} + b_{j \oplus 2} = d$ we obtain $b_j = b_{j \oplus 2}$. Since T is odd by lemma A2, it now follows that $b_0 = b_1 = \ldots = b_{T-1}$, and by using Eq. (A26) we obtain

$$b_0 = b_1 = \ldots = b_{T-1} = d/2. \tag{A27}$$

From expression (A20) and Eq. (A27) we obtain $\mathcal{S}(h) = \{d/2\}$. Therefore, for $x \in Z^2$,

$$\begin{aligned} f(x) &= g * h(x) \\ &= h(d/2) g(x - d/2). \end{aligned} \tag{A28}$$

If $h(d/2) = 0$, then f would be identically zero, contradicting the assumption that $R(M) = \mathcal{S}(f)$. Therefore $h(d/2) \neq 0$. For $x \in Z^2$,

$$\begin{aligned} f_1(x) &= g * h_1(x - d) \\ &= h(d/2)^* g(x - d + d/2) \\ &= h(d/2)^* g(x - d/2) \\ &= \alpha f(x), \end{aligned} \tag{A29}$$

where

$$\alpha = \frac{h(d/2)^*}{h(d/2)}.$$

Since $|\alpha| = 1$, this completes the proof of the theorem.

APPENDIX B

It will be shown in this appendix that the program presented at the end of Section 4 computes $f(q_n)$ for $n = T, \ldots, N-1$. It will be assumed that $f(q_n)$ for $n = 0, \ldots, T-1$ has been computed as described in Section 4. Since $0 \leq m_n \leq T-1$, $f(m_n)$ has been computed for $n = T, \ldots, N-1$.

For $T \leq n \leq N-1$ we have

$$r(q_n - q_{m_n}) = \sum_{y \in Z^2} f(y) f(y - q_n + q_{m_n})^*. \tag{B1}$$

If $y \in \mathcal{S}(f)$ and $y = q_n + q_{m_n} \in \mathcal{S}(f)$, then, since $\mathcal{S}(f) \subseteq M$, it follows that $y \in M$ and $y = q_n + q_{m_n} \in M$ or, equivalently, $y \in M + q_n - q_{m_n}$. Therefore

$$y \in M \cap (M + q_n - q_{m_n}) \subseteq \{q_0, \ldots, q_{n}\}. \tag{B2}$$

Hence

$$\begin{aligned} r(q_n - q_{m_n}) &= \sum_{k=0}^{n} f(q_k) f(q_k - q_n + q_{m_n})^* \\ &= f(q_n) f(q_{m_n})^* + \sum_{k=0}^{n-1} f(q_k) f(q_k - q_n + q_{m_n})^*. \end{aligned} \tag{B3}$$

Thus

$$f(q_n) = \frac{1}{f(q_{m_n})^*} \left[r(q_n - q_{m_n}) - \sum_{k=0}^{n-1} f(q_k) f(q_k - q_n + q_{m_n})^* \right]. \tag{B4}$$

An induction argument will be used to prove that $f(q_n)$ is computed correctly for $n = T, \ldots, N-1$. The induction hypothesis is

$$H(n): \quad f(q_j) \text{ is computed correctly for } 0 \le j \le n. \quad \text{(B5)}$$

By the derivation in Section 4, $H(T-1)$ is true. Now assume that $T \le n \le N-1$ and $H(n-1)$ is true. We want to show that $H(n)$ is true. It suffices to show that $f(q_n)$ is computed correctly. Let all variables have the values that they have at step 2 of the pass through the loop in which $f(q_n)$ is computed. It must be shown that all values of f appearing in the right-hand side of Eq. (B4) are correct. Since, by assumption, $H(n-1)$ is true, it follows that $f(q_k)$, $k = 0, \ldots, n-1$ have the correct values. Also, as mentioned above, $f(q_{m_n})$ has the correct value. Let $x = q_k - q_n + q_{m_n}$ with $0 \le k \le n-1$. If $x \notin M$, then $f(x) = 0$. In this case, since $\mathcal{S}(f) \subseteq M$, it follows that $x \notin \mathcal{S}(f)$, and therefore 0 is the correct value for $f(x)$. Assume that $x \in M$. We have $x + q_n - q_{m_n} = q_k \in M$ and therefore $x \in M - q_n + q_{m_n}$. Hence $x \in M \cap (M - q_n + q_{m_n}) \subseteq \{q_0, \ldots, q_{n-1}\}$. It now follows from the induction hypothesis, $H(n-1)$, that $f(x)$ has the correct value when the value for $f(q_n)$ is computed. Therefore $f(q_n)$ is computed correctly, and $H(n)$ is true. By induction it follows that $H(N-1)$ is true and hence $f(n)$ is computed correctly for $n = 0, \ldots, N-1$.

APPENDIX C

In this appendix it will be shown that the program in Section 5 generates a reconstruction algorithm. First, it will be shown that the loop is not infinite, and hence the program produces sequences $q_T, \ldots, q_{N-1}$ and $m_T, \ldots, m_{N-1}$. Second, it will be shown that if $q = (q_0, \ldots, q_{N-1})$, and $m = (m_T, \ldots, m_{N-1})$, then (q, m) is a reconstruction algorithm.

In what follows, 0 will be used to denote both the number zero and the origin of $\mathcal{R}^2$. Context should prevent any confusion.

If $x, y, z \in \mathcal{R}^2$, let $[x, y, z]$ denote their convex hull in $\mathcal{R}^2$. If x, y and z are noncollinear, then the interior of $[x, y, z]$ is given by

$$\text{int}[x, y, z] = \{ax + by + cz : a, b, c > 0 \text{ and } a + b + c = 1\}; \quad \text{(C1)}$$

then $0 \in \text{int}[x, y, z]$ if and only if x, y, and z are noncollinear and there exist strictly positive numbers a, b, c such that $ax + by + cz = 0$. (The sum of a, b, and c can be made equal to 1 by dividing each of these numbers by their sum if necessary.)

The following lemma will be needed.

Lemma C1

If $\mu_n, \nu_n \in \mathcal{R}^2$, $n = 1, 2, 3$, $\langle \mu_n, \nu_n \rangle > 0$, and $\langle \mu_n, \nu_m \rangle < 0$ for $n \ne m$, then $0 \in \text{int}[\mu_1, \mu_2, \mu_3]$ and $0 \in \text{int}[\nu_1, \nu_2, \nu_3]$.

Proof

By symmetry it suffices to show that $0 \in \text{int}[\mu_1, \mu_2, \mu_3]$.

First, we will show that μ_1, μ_2, and μ_3 are noncollinear. If they were collinear, then one of them would be in the convex hull of the other two. If μ_1 were in the convex hull of μ_2 and μ_3, then there would exist numbers τ_2 and τ_3 such that $\tau_2, \tau_3 \ge 0$; $\tau_2 + \tau_3 = 1$; and $\mu_1 = \tau_2\mu_2 + \tau_3\mu_3$. In that case, however, $\langle \mu_1, \nu_1 \rangle = \langle \tau_2\mu_2 + \tau_3\mu_3, \nu_1 \rangle = \tau_2\langle \mu_2, \nu_1 \rangle + \tau_3\langle \mu_3, \nu_1 \rangle < 0$, contradicting the assumption that $\langle \mu_1, \nu_1 \rangle > 0$. Therefore μ_1, μ_2, and μ_3 are noncollinear.

Since any three vectors in $\mathcal{R}^2$ are linearly dependent, there exist three numbers σ_1, σ_2, and σ_3, not all zero, such that

$$\sigma_1\mu_1 + \sigma_2\mu_2 + \sigma_3\mu_3 = 0. \quad \text{(C2)}$$

Since $\langle \mu_n, \nu_n \rangle > 0$, $\mu_n \ne 0$, $n = 1, 2, 3$. Therefore no two of the σ_n can be zero.

Now suppose that $\sigma_1 = 0$; then $\sigma_2 \ne 0 \ne \sigma_3$, and it follows from Eq. (C2) that

$$\mu_2 = -(\sigma_3/\sigma_2)\mu_3; \quad \text{(C3)}$$

hence

$$0 < \langle \mu_2, \nu_2 \rangle = -(\sigma_3/\sigma_2)\langle \mu_3, \nu_2 \rangle. \quad \text{(C4)}$$

Since $\langle \mu_3, \nu_2 \rangle < 0$, it follows from expression (C4) that

$$-\sigma_3/\sigma_2 < 0. \quad \text{(C5)}$$

Also,

$$0 > \langle \mu_2, \nu_1 \rangle = -(\sigma_3/\sigma_2)\langle \mu_3, \nu_1 \rangle, \quad \text{(C6)}$$

and since $\langle \mu_3, \nu_1 \rangle < 0$, it follows from expression (C6) that $-\sigma_3/\sigma_2 > 0$, which contradicts expression (C5). Therefore $\sigma_1 \ne 0$, and by symmetry, $\sigma_2 \ne 0 \ne \sigma_3$. By multiplying the σ_n's by -1 if necessary, we may assume that $\sigma_1 > 0$. We have

$$\begin{aligned} \sigma_1\langle \mu_1, \nu_1 \rangle + \sigma_2\langle \mu_2, \nu_1 \rangle + \sigma_3\langle \mu_3, \nu_1 \rangle &= \langle \sigma_1\mu_1 + \sigma_2\mu_2 + \sigma_3\mu_3, \nu_1 \rangle \\ &= 0. \end{aligned} \quad \text{(C7)}$$

Since $\sigma_1 > 0$ and $\langle \mu_1, \nu_1 \rangle > 0$,

$$\sigma_2\langle \mu_2, \nu_1 \rangle + \sigma_3\langle \mu_3, \nu_1 \rangle = -\sigma_1\langle \mu_1, \nu_1 \rangle < 0. \quad \text{(C8)}$$

Since $\langle \mu_2, \nu_1 \rangle < 0$ and $\langle \mu_3, \nu_1 \rangle < 0$, at least one of the numbers σ_2 and σ_3 must be strictly positive. By symmetry, we may assume without loss of generality that $\sigma_2 > 0$. Now

$$\begin{aligned} \sigma_1\langle \mu_1, \nu_3 \rangle + \sigma_2\langle \mu_2, \nu_3 \rangle + \sigma_3\langle \mu_3, \nu_3 \rangle &= \langle \sigma_1\mu_1 + \sigma_2\mu_2 + \sigma_3\mu_3, \nu_3 \rangle \\ &= 0. \end{aligned} \quad \text{(C9)}$$

Since $\sigma_1 > 0$, $\sigma_2 > 0$, $\langle \mu_1, \nu_3 \rangle < 0$, and $\langle \mu_2, \nu_3 \rangle < 0$, it follows that

$$\begin{aligned} \sigma_3\langle \mu_3, \nu_3 \rangle &= -\sigma_1\langle \mu_1, \nu_3 \rangle - \sigma_2\langle \mu_2, \nu_3 \rangle \\ &> 0. \end{aligned} \quad \text{(C10)}$$

Since $\langle \mu_3, \nu_3 \rangle > 0$, it follows that $\sigma_3 > 0$. By the comment preceding the lemma, it now follows that $0 \in \text{int}[\mu_1, \mu_2, \mu_3]$. This completes the proof of lemma C1.

One more lemma is needed before proving that the loop is not infinite. As in Appendix A, we define $j \oplus k = (j + k) \bmod T$.

Lemma C2

For $j = 0, \ldots, T-1$ and $k = 2, \ldots, T-1$, $0 \in \text{int}[y_j, y_{j\oplus 1}, (-1)^k y_{j\oplus k}]$.

Proof

The proof will be by induction on k. Let $k = 2$. We want to show that $0 \in \text{int}[y_j, y_{j\oplus 1}, y_{j\oplus 2}]$. Let

$$\mu_1 = q_{j\oplus 1} - q_{j\oplus 3}, \quad \mu_2 = q_{j\oplus 2} - q_{j\oplus 1}, \quad \mu_3 = q_j - q_{j\oplus 2},$$

$$\nu_1 = y_j, \quad \nu_2 = y_{j\oplus 1}, \quad \nu_3 = y_{j\oplus 2}. \tag{C11}$$

By lemma A3, $\langle \mu_n, \nu_n \rangle > 0$ and $\langle \mu_n, \nu_m \rangle < 0$ for $n \neq m$. Therefore, by lemma C1, $0 \in \text{int}[\nu_1, \nu_2, \nu_3] = \text{int}[y_j, y_{j\oplus 1}, y_{j\oplus 2}]$.

Let $3 \leq k \leq T-1$, and assume that the lemma is true for $k-1$. We want to show that $0 \in \text{int}[y_j, y_{j\oplus 1}, (-1)^k y_{j\oplus k}]$. We have shown that $0 \in \text{int}[y_j, y_{j\oplus 1}, y_{j\oplus 2}]$, and therefore there exist strictly positive numbers $\sigma_1, \sigma_2, \sigma_3$ such that

$$\sigma_1 y_j + \sigma_2 y_{j\oplus 1} + \sigma_3 y_{j\oplus 2} = 0. \tag{C12}$$

Applying the lemma for $k-1$ with j replaced by $j \oplus 1$, we have $0 \in \text{int}[y_{j\oplus 1}, y_{j\oplus 2}, (-1)^{k-1} y_{j\oplus k}]$, and therefore there exist strictly positive numbers τ_1, τ_2, τ_3 such that

$$\tau_1 y_{j\oplus 1} + \tau_2 y_{j\oplus 2} + \tau_3 (-1)^{k-1} y_{j\oplus k} = 0. \tag{C13}$$

By multiplying Eq. (C13) by σ_3/τ_2 and subtracting the result from Eq. (C12), we obtain

$$\lambda_1 y_j + \lambda_2 y_{j\oplus 1} + \lambda_3 (-1)^k y_{j\oplus k} = 0, \tag{C14}$$

where $\lambda_1 = \sigma_1$, $\lambda_2 = \sigma_2 - \sigma_3 \tau_1/\tau_2$, and $\lambda_3 = \sigma_3 \tau_3/\tau_2$. We have $\lambda_1 > 0$ and $\lambda_3 > 0$. We will show that $\lambda_2 > 0$. From Eq. (C14) we obtain

$$\lambda_2 y_{j\oplus 1} = -\lambda_1 y_j + \lambda_3 (-1)^{k-1} y_{j\oplus k}. \tag{C15}$$

We consider two cases.

Case 1
k is odd; then $k-1$ is even, and, by Eq. (C15),

$$\lambda_2 y_{j\oplus 1} = -\lambda_1 y_j + \lambda_3 y_{j\oplus k}; \tag{C16}$$

and, by using lemma A3,

$$\begin{aligned} &\lambda_2 \langle q_{j\oplus k\oplus 1} - q_{j\oplus 1}, y_{j\oplus 1} \rangle \\ &\quad = -\lambda_1 \langle q_{j\oplus k\oplus 1} - q_{j\oplus 1}, y_j \rangle + \lambda_3 \langle q_{j\oplus k\oplus 1} - q_{j\oplus 1}, y_{j\oplus k} \rangle \\ &\quad = \lambda_1 \langle q_{j\oplus 1} - q_{j\oplus k\oplus 1}, y_j \rangle + \lambda_3 \langle q_{j\oplus k\oplus 1} - q_{j\oplus 1}, y_{j\oplus k} \rangle \\ &\quad > 0. \end{aligned} \tag{C17}$$

Since, by lemma A3, $\langle q_{j\oplus k\oplus 1} - q_{j\oplus 1}, y_{j\oplus 1} \rangle > 0$, it follows from expression (C17) that $\lambda_2 > 0$.

Case 2
k is even; then $k-1$ is odd, and, by Eq. (C15),

$$\lambda_2 y_{j\oplus 1} = -\lambda_1 y_j - \lambda_3 y_{j\oplus k}; \tag{C18}$$

and, by using lemma A3,

$$\begin{aligned} \lambda_2 \langle q_{j\oplus k} - q_{j\oplus 1}, y_{j\oplus 1} \rangle &= \lambda_1 \langle q_{j\oplus 1} - q_{j\oplus k}, y_j \rangle \\ &\quad + \lambda_3 \langle q_{j\oplus 1} - q_{j\oplus k}, y_{j\oplus k} \rangle \\ &> 0. \end{aligned} \tag{C19}$$

Since, by lemma A3, $\langle q_{j\oplus k} - q_{j\oplus 1}, y_{j\oplus 1} \rangle > 0$, it follows from expression (C19) that $\lambda_2 > 0$.

It remains to be shown that y_j, $y_{j\oplus 1}$, and $(-1)^k y_{j\oplus k}$ are noncollinear. Since $\lambda_1, \lambda_2, \lambda_3 > 0$, it follows from Eq. (C14) that $0 \in [y_j, y_{j\oplus 1}, (-1)^k y_{j\oplus k}]$. Therefore, if y_j, $y_{j\oplus 1}$, and $(-1)^k y_{j\oplus k}$ are collinear, then they must all lie on a line through the origin. However, since we have already shown that $0 \in \text{int}[y_j, y_{j\oplus 1}, y_{j\oplus 2}]$, y_j and $y_{j\oplus 1}$ cannot lie on a line through the origin. Therefore y_j, $y_{j\oplus 1}$, and $(-1)^k y_{j\oplus k}$ are noncollinear, and hence $0 \in \text{int}[y_j, y_{j\oplus 1}, (-1)^k y_{j\oplus k}]$. This completes that proof of lemma C2.

In order to prove that the loop is not infinite, it will be shown that the parameter n in the program in Section 5 can fail to be incremented on at most $T-2$ consecutive passes through the loop. The proof will be by contradiction. Accordingly, assume that n is not incremented on $T-1$ consecutive passes through the loop.

Let k and ϕ have the values that they have at step 2 of the first of these $T-1$ passes. Let

$$b_a = \min\{j: 0 \leq j \leq N-T-1 \text{ and } \phi(d_{k\oplus a,j}) = 1\} \tag{C20}$$

for $a = 0, \ldots, T-2$; then

$$\phi(d_{k\oplus a,b_a}) = 1 \tag{C21}$$

and

$$\phi(\alpha_{k\oplus a} - d_{k\oplus a,b_a}) = 1 \tag{C22}$$

for $a = 0, \ldots, T-2$. Let

$$x_a = \begin{cases} d_{k\oplus a,b_a} & \text{if } h_{k\oplus a}(d_{k\oplus a,b_a}) \geq 0 \\ \alpha_{k\oplus a} - d_{k\oplus a,b_a} & \text{otherwise} \end{cases} \tag{C23}$$

We have

$$h_{k\oplus a}(x_a) = |h_{k\oplus a}(d_{k\oplus a,b_a})|. \tag{C24}$$

If $x \in Z^2$ and $\phi(x) = 1$, then $x \in D$ and $x = d_{k\oplus a,j}$ for some $j \geq b_a$. Therefore $|h_{k\oplus a}(x)| = |h_{k\oplus a}(d_{k\oplus a,j})| \leq |h_{k\oplus a}(d_{k\oplus a,b_a})| = h_{k\oplus a}(x_a)$. Thus, for $x \in Z^2$,

$$\phi(x) = 1 \Rightarrow |h_{k\oplus a}(x)| \leq h_{k\oplus a}(x_a). \tag{C25}$$

Claim C1
For $a = 0, \ldots, T-2$,

$$(-1)^a \langle x_a - q_{k\oplus a\oplus 1}, y_{k\oplus (T-1)} \rangle > 0. \tag{C26}$$

Proof of Claim
First, we will prove the claim for $a = 0$. Since by Eqs. (C21)–(C23), $\phi(\alpha_k - x_0) = 1$, it follows that $\alpha_k - x_0 \in D$. Therefore, by lemma A3,

$$\langle \alpha_k - x_0, y_{k\oplus (T-1)} \rangle < \langle q_k, y_{k\oplus (T-1)} \rangle, \tag{C27}$$

or, since $\alpha_k = q_k + q_{k\oplus 1}$,

$$\langle x_0 - q_{k\oplus 1}, y_{k\oplus (T-1)} \rangle > 0. \tag{C28}$$

Thus the claim is true for $a = 0$. Now let $1 \leq a \leq T-2$ and assume the claim is true for $a-1$. By Eqs. (C21)–(C23), $\phi(\alpha_{k\oplus a} - x_a) = 1$, and hence, by using implication (C25),

$$\begin{aligned} h_{k\oplus (a-1)}(\alpha_{k\oplus a} - x_a) &\leq |h_{k\oplus (a-1)}(\alpha_{k\oplus a} - x_a)| \\ &\leq h_{k\oplus (a-1)}(x_{a-1}), \end{aligned} \tag{C29}$$

or, equivalently,

$$\langle x_{a-1} - \alpha_{k\oplus a} + x_a, y_{k\oplus (a-1)} \rangle \geq 0. \tag{C30}$$

Similarly, $\phi(x_{a-1}) = 1$, and

$$-h_{k\oplus a}(x_{a-1}) \le |h_{k\oplus a}(x_{a-1})| \le h_{k\oplus a}(x_a) = -h_{k\oplus a}(\alpha_{k\oplus a} - x_a), \quad \text{(C31)}$$

and therefore

$$h_{k\oplus a}(\alpha_{k\oplus a} - x_a) \le h_{k\oplus a}(x_{a-1}), \quad \text{(C32)}$$

or, equivalently,

$$\langle x_{a-1} - \alpha_{k\oplus a} + x_a, y_{k\oplus a}\rangle \ge 0. \quad \text{(C33)}$$

By lemma C2, $0 \in \text{int}[y_{k\oplus(a-1)}, y_{k\oplus a}, (-1)^{T-a}y_{k\oplus(T-1)}]$, and therefore there exist strictly positive real numbers $\sigma_1, \sigma_2, \sigma_3$ such that

$$\sigma_1 y_{k\oplus(a-1)} + \sigma_2 y_{k\oplus a} + \sigma_3(-1)^{T-a}y_{k\oplus(T-1)} = 0. \quad \text{(C34)}$$

Since, by lemma A2, T is odd, $(-1)^{T-a} = -(-1)^a$, and from Eq. (C34) we obtain

$$\sigma_3(-1)^a y_{k\oplus(T-1)} = \sigma_1 y_{k\oplus(a-1)} + \sigma_2 y_{k\oplus a}. \quad \text{(C35)}$$

By Eq. (C35) and expressions (C30) and (C33),

$$\sigma_3(-1)^a\langle x_{a-1} - \alpha_{k\oplus a} + x_a, y_{k\oplus(T-1)}\rangle = \sigma_1\langle x_{a-1} - \alpha_{k\oplus a} + x_a, y_{k\oplus(a-1)}\rangle + \sigma_2\langle x_{a-1} - \alpha_{k\oplus a} + x_a, y_{k\oplus a}\rangle \ge 0, \quad \text{(C36)}$$

and since $\sigma_3 > 0$,

$$(-1)^a\langle x_{a-1} - \alpha_{k\oplus a} + x_a, y_{k\oplus(T-1)}\rangle \ge 0. \quad \text{(C37)}$$

By substituting $\alpha_{k\oplus a} = q_{k\oplus a} + q_{k\oplus a\oplus 1}$ and using expression (C37) and the induction hypothesis, we obtain

$$(-1)^a\langle x_a - q_{k\oplus a\oplus 1}, y_{k\oplus(T-1)}\rangle \ge -(-1)^a\langle x_{a-1} - q_{k\oplus a}, y_{k\oplus(T-1)}\rangle = (-1)^{a-1}\langle x_{a-1} - q_{k\oplus a}, y_{k\oplus(T-1)}\rangle > 0. \quad \text{(C38)}$$

This completes the proof of the claim.

Set $a = T - 2$ in expression (C26). Since, by lemma A2, $T - 2$ is odd, we obtain

$$\langle x_{T-2}, y_{k\oplus(T-1)}\rangle < \langle q_{k\oplus(T-1)}, y_{k\oplus(T-1)}\rangle. \quad \text{(C39)}$$

It now follows from lemma A3 that $x_{T-2} \notin D$, and therefore $\phi(x_{T-2}) = 0$, which contradicts either Eq. (C21) or Eq. (C22). Therefore n cannot fail to be incremented on each of $T - 1$ consecutive passes through the loop. This completes the proof that the loop is not infinite.

It now follows that the program produces sequences $q_T, \ldots, q_{N-1}$ and $m_T, \ldots, m_{N-1}$. It remains to be shown that if $q = (q_0, \ldots, q_{N-1})$ and $m = (m_T, \ldots, m_{N-1})$, then (q, m) is a reconstruction algorithm for the mask M.

We have $R(M) = |q_0, \ldots, q_{T-1}|$. Let $T \le n \le N - 1$, and let all variables have the values that they have after step 5 and before step 6 of the pass through loop in which q_n and m_n are defined. By the definition of b in step 1 of the loop, for $x \in Z^2$,

$$|h_k(d_{k,b})| < |h_k(x)| \Rightarrow \phi(x) = 0. \quad \text{(C40)}$$

If $x \in D$, then before entering the loop ϕ had to have the value 1 at x. Thus, if the current value is $\phi(x) = 0$, ϕ must have acquired the value 0 at x on some preceding pass; that is, $x = q_{n'}$ for some $n' < n$. Therefore $x \in |q_0, \ldots, q_{n-1}|$. Thus, for $x \in D$,

$$\phi(x) = 0 \Rightarrow x \in |q_0, \ldots, q_{n-1}|. \quad \text{(C41)}$$

First, it will be shown that

$$M \cap (M + q_n - q_{m_n}) \subseteq |q_0, \ldots, q_n|. \quad \text{(C42)}$$

Let $x \in M \cap (M + q_n - q_{m_n})$. We want to show that $x \in |q_0, \ldots, q_n|$. If $x \in R(M)$, then, since $n \ge T$, we are done. Since $q_n = d_{k,b}$, if $x = d_{k,b}$ we are done. Now assume that $x \notin R(M)$ and $x \ne d_{k,b}$. Since $x \notin R(M)$, $x \in D$.

Claim C2

$$|h_k(d_{k,b})| < |h_k(x)|. \quad \text{(C43)}$$

Proof of Claim

The proof of the claim will be divided into two cases.

Case 1

$$h_k(d_{k,b}) \ge 0. \quad \text{(C44)}$$

Since $q_n = d_{k,b}$, it follows from step 5 of the loop that $m_n = k$. Let $z = x' - d_{k,b} + q_k$. Since $x \in M + d_{k,b} - q_k$, it follows that $z \in M$. Also, since $x \ne d_{k,b}$, $z \ne q_k$. Therefore, by lemma A3,

$$\langle q_k, y_k\rangle < \langle z, y_k\rangle = \langle x, y_k\rangle - \langle d_{k,b}, y_k\rangle + \langle q_k, y_k\rangle, \quad \text{(C45)}$$

or $\langle d_{k,b}, y_k\rangle < \langle x, y_k\rangle$. It follows that $h_k(d_{k,b}) < h_k(x)$, and since $h_k(d_{k,b}) \ge 0$, the claim follows.

Case 2

$$h_k(d_{k,b}) < 0. \quad \text{(C46)}$$

In this case $m_n = k \oplus 1$. Let $z = x - d_{k,b} + q_{k\oplus 1}$. Since $x \in M + d_{k,b} - q_{k\oplus 1}$, it follows that $z \in M$. Also, since $x \ne d_{k,b}$, $z \ne q_{k\oplus 1}$. Therefore, by lemma A3,

$$\langle x, y_k\rangle - \langle d_{k,b}, y_k\rangle + \langle q_{k\oplus 1}, y_k\rangle = \langle z, y_k\rangle < \langle q_{k\oplus 1}, y_k\rangle, \quad \text{(C47)}$$

or $\langle x, y_k\rangle < \langle d_{k,b}, y_k\rangle$. It follows that $h_k(x) < h_k(d_{k,b})$, and since $h_k(d_{k,b}) < 0$, the claim follows. This completes the proof of the claim.

It follows by implication (C40) that $\phi(x) = 0$, and since $x \in D$, it follows from implication (C41) that $x \in |q_0, \ldots, q_{n-1}| \subseteq |q_0, \ldots, q_n|$. This completes the proof of relation (C42).

It remains to be shown that

$$M \cap (M - d_{k,b} + q_{m_n}) \subseteq |q_0, \ldots, q_{n-1}|. \quad \text{(C48)}$$

Let $x \in M \cap (M - d_{k,b} + q_{m_n})$. We want to show that $x \in |q_0, \ldots, q_{n-1}|$. Since $n \ge T$, if $x \in R(M)$, we are done. Assume that $x \notin R(M)$; then $x \in D$. The proof will be divided into two cases.

Case 1

$h_k(d_{k,b}) \geq 0$.

In this case $m_n = k$. Let $z = x + d_{k,b} - q_k$. Since $x \in M - d_{k,b} + q_k$, it follows that $z \in M$.

If $z = q_{k \oplus 1}$, then $x = q_k + q_{k \oplus 1} - d_{k,b} = \alpha_k - d_{k,b}$. If $\phi(x) = 1$, then by step 2 of the loop, q_n would not have been defined on this pass, contrary to the assumption that it was. Therefore $\phi(x) = 0$, and by implication (C41), $x \in |q_0, \ldots, q_{n-1}|$.

Assume that $z \neq q_{k \oplus 1}$; then, by lemma A3,

$$\langle x, y_k \rangle + \langle d_{k,b}, y_k \rangle - \langle q_k, y_k \rangle = \langle z, y_k \rangle < \langle q_{k \oplus 1}, y_k \rangle. \quad \text{(C49)}$$

Recalling that $\beta_k = (q_k + q_{k \oplus 1})/2$, it follows from expression (C49) that

$$\langle d_{k,b} - \beta_k, y_k \rangle < \langle -x + \beta_k, y_k \rangle, \quad \text{(C50)}$$

or $h_k(d_{k,b}) < -h_k(x)$, and since $h_k(d_{k,b}) \geq 0$, it follows that $|h_k(d_{k,b})| < |h_k(x)|$. Now by implication (C40), $\phi(x) = 0$, and by implication (C41), $x \in |q_0, \ldots, q_{n-1}|$.

Case 2

$h_k(d_{k,b}) < 0$.

In this case $m_n = k \oplus 1$. Let $z = x + d_{k,b} - q_{k \oplus 1}$. Since $x \in M - d_{k,b} + q_{k \oplus 1}$, it follows that $z \in M$.

If $z = q_k$, then $x = q_k + q_{k \oplus 1} - d_{k,b} = \alpha_k - d_{k,b}$, and by the argument given in case 1 above, $\phi(x) = 0$, and by implication (C41), $x \in |q_0, \ldots, q_{n-1}|$.

Assume that $z \neq q_k$; then, by lemma A3,

$$\langle q_k, y_k \rangle < \langle z, y_k \rangle = \langle x, y_k \rangle + \langle d_{k,b}, y_k \rangle - \langle q_{k \oplus 1}, y_k \rangle, \quad \text{(C51)}$$

from which it follows that $-h_k(x) < h_k(d_{k,b})$. Since $h_k(d_{k,b}) < 0$, it follows that $|h_k(d_{k,b})| < |h_k(x)|$. Hence, by implication (C40), $\phi(x) = 0$, and by implication (C41), $x \in |q_0, \ldots, q_{n-1}|$. This establishes relation (C48) and completes the proof that (q, m) is a reconstruction algorithm for the mask M.

ACKNOWLEDGMENT

This research was supported by Air Force Wright Aeronautical Laboratories, Avionics Laboratory, under contract F33615-83-C-1046, DARPA order 5205.

REFERENCES

1. J. R. Fienup, T. R. Crimmins, and W. Holsztynski, "Reconstruction of the support of an object from the support of its autocorrelation function," J. Opt. Soc. Am. 72, 610–624 (1982).
2. M. H. Hayes and T. F. Quatieri, "Recursive phase retrieval using boundary conditions," J. Opt. Soc. Am. 73, 1427–1433 (1983).
3. Y. M. Bruck and L. G. Sodin, "On the ambiguity of the image reconstruction problem," Opt. Commun. 30, 304–308 (1979).
4. M. A. Fiddy, B. J. Brames, and J. C. Dainty, "Enforcing irreducibility for phase retrieval in two dimensions," Opt. Lett. 8, 96–98 (1983).
5. J. R. Fienup, "Phase retrieval algorithms: a comparison," Appl. Opt. 21, 2758–2769 (1982).
6. J. R. Fienup, "Reconstruction of an object from the modulus of its Fourier transform," Opt. Lett. 3, 27–29 (1978).
7. J. R. Fienup, "Space object imaging through the turbulent atmosphere," Opt. Eng. 18, 529–534 (1979).
8. J. R. Fienup, "Reconstruction of objects having latent reference points," J. Opt. Soc. Am. 73, 1421–1426 (1983).
9. B. J. Brames, "Unique phase retrieval with explicit support information," Opt. Lett. 11, 61–63 (1986).
10. J. W. Goodman, "Analogy between holography and interferometric image formation," J. Opt. Soc. Am. 60, 506–509 (1970).
11. C. Y. C. Liu and A. W. Lohman, "High resolution image formation through the turbulent atmosphere," Opt. Commun. 8, 372–377 (1973).

Appendix E
$\nabla e_s^2(g(x))$ FOR COMPLEX OBJECTS

In this appendix we generalize the expression for the gradient of the summed objective function to include objects and object estimates that are complex valued.

Recall that the summed objective function is defined as the sum of a generalized object-domain error metric and the Fourier-domain error metric:

$$e_s^2 = \epsilon_0^2 + e_F^2, \tag{E-1}$$

where

$$\epsilon_0^2 = \sum_{x \in S'} |g(x)|^2, \tag{E-2}$$

and

$$e_F^2 = N^{-2} \sum_u [\,|G(u)| - |F(u)|\,]^2. \tag{E-3}$$

Notice the summed objective function is implicitly a function of the real and imaginary parts of the pixel values for the latest estimate, $g(x)$. We therefore treat the real and imaginary parts of each pixel as distinct parameters that can be adjusted in order to minimize e_s^2. We express the real and imaginary parts of the latest estimate as

$$g(x) = a(x) + ib(x). \tag{E-4}$$

We define the gradients to be:

$$\nabla e_s^2(g(x)) = \sum_{j=1}^{N^2} \frac{\partial e_s^2}{\partial a(x_j)} v_j^R + \frac{\partial e_s^2}{\partial b(x_j)} v_j^I \tag{E-5}$$

where v_j^R and v_j^I are orthogonal unit vectors associated with the real and imaginary parts of the j^{th} pixel. The first partial derivative in Eq. (E-5) may be separated into two terms:

$$\frac{\partial e_s^2}{\partial a(x_j)} = \frac{\partial e_F^2}{\partial a(x_j)} + \frac{\partial e_o^2}{\partial a(x_j)} \quad . \tag{E-6}$$

For the moment we examine the first term in (E-6):

$$\frac{\partial e_F^2}{\partial a(x_j)} = \frac{\partial}{\partial a(x_j)} N^{-2} \sum_u \left[|G(u)| - |F(u)|\right]^2 \tag{E-7}$$

$$= 2N^{-2} \sum_u (|G(u)| - |F(u)|) \frac{\partial |G(u)|}{\partial a(x_j)} \quad . \tag{E-8}$$

It is easy to see that

$$\frac{\partial |G(u)|}{\partial a(x_j)} = \frac{1}{2|G(u)|} \frac{\partial |G(u)|^2}{\partial a(x_j)}$$

$$= \frac{1}{2|G(u)|} \left[G(u)\, e^{i2\pi u \cdot x_j/N} + C.C.\right] \quad . \tag{E-9}$$

Substituting Eq. (E-9) into Eq. (E-8) yields

$$\frac{\partial e_F^2}{\partial a(x_j)} = N^{-2} \sum_u (|G(u)| - |F(u)|) \frac{(G(u)\, e^{i2\pi u \cdot x_j / N} + C.C.)}{|G(u)|} \qquad (E\text{-}10)$$

$$= (g(x_j) - g'(x_j)) + (g^*(x_j) - g'^*(x_j))$$

$$= 2(a(x_j) - a'(x_j)).$$

This result is consistent with the result quoted for real-valued objects. A parallel derivation gives

$$\frac{\partial e_F^2}{\partial b(x_j)} = 2(b(x_j) - b'(x_j)) \quad . \qquad (E\text{-}13)$$

We now return to the second term in Eq. (E-6):

$$\frac{\partial \varepsilon_o^2}{\partial a(x_j)} = \frac{\partial}{\partial a(x_j)} \sum_{x \in S'} |g(x)|^2$$

$$= \frac{\partial}{\partial a(x_j)} \sum_{x \in S'} a^2(x) + b^2(x)$$

$$= \begin{cases} 2a(x_j) & , x_j \in S' \\ 0 & , x_j \in S \end{cases} \quad . \qquad (E\text{-}14)$$

Similarly,

$$\frac{\partial \varepsilon_o^2}{\partial b(x_j)} = \begin{cases} 2b(x_j) & , x_j \in S' \\ 0 & , x_j \in S \end{cases} \qquad (E\text{-}15)$$

Let

$$S'(x) = \begin{cases} 1, & x \in S' \\ 0, & x \in S \end{cases} \qquad \text{(E-16)}$$

Now Eqs. (E-14) and (E-15) may be expressed more conveniently:

$$\frac{\partial \epsilon_o^2}{\partial a(x_j)} = 2a(x_j)S'(x_j) \qquad \text{(E-17)}$$

and

$$\frac{\partial \epsilon_o^2}{\partial b(x_j)} = 2b(x_j)S'(x_j) \quad . \qquad \text{(E-18)}$$

Collecting these results, we have:

$$\frac{\partial e_s^2}{\partial a(x_j)} = 2[a(x_j) - a'(x_j)] + 2a(x_j)S'(x_j) \qquad \text{(E-19)}$$

$$\frac{\partial e_s^2}{\partial b(x_j)} = 2[b(x_j) - b'(x_j)] + 2b(x_j)S'(x_j) \qquad \text{(E-20)}$$

Equations (E-19) and (E-20) may be combined to form a complex gradient image.

$$\text{Gradient image} \equiv 2[g(x) - g'(x)] + 2g(x)S'(x) \qquad \text{(E-21)}$$

The extension to complex-valued objects still requires only 2 FFTs to compute the gradient.

Appendix F
$\nabla e_o^2(g(x))$ FOR REAL OBJECTS

Recall that the object-domain error metric is implicitly a function of the input estimate $g(x)$ and is given by:

$$e_o^2(g(x)) = \sum_{x \in S'} [g'(x)]^2. \tag{F-1}$$

Its gradient with respect to the input pixel values may be written:

$$\nabla e_o^2(g(x)) = \sum_{j=1}^{N^2} \frac{\partial e_o^2}{\partial g(x_j)} \, v_j \tag{F-2}$$

where v_j is a unit vector in the direction of the parameter $g(x_j)$ in parameter space. We focus now on the partial derivative that appears in Eq. (F-2):

$$\frac{\partial e_o^2}{\partial g(x_j)} = \frac{\partial}{\partial g(x_j)} \sum_{x \in S'} \left[g'(x) \right]^2$$

$$= 2 \sum_{x \in S'} g'(x) \frac{\partial}{\partial g(x_j)} g'(x) \, . \tag{F-3}$$

Substitution of the Fourier-domain expression for $g'(x)$ gives:

$$\frac{\partial e_o^2}{\partial g(x_j)} = 2 \sum_{x \in S'} g'(x) \frac{\partial}{\partial g(x_j)} \left\{ N^{-2} \sum_u G'(u) \, e^{i 2\pi u \cdot x / N} \right\} \, . \tag{F-4}$$

Recall that

$$G'(u) = \frac{G(u)|F(u)|}{|G(u)|} \tag{F-5}$$

giving

$$\frac{\partial e_o^2}{\partial g(x_j)} = 2 \sum_{x \in S'} g'(x) \frac{\partial}{\partial g(x_j)} \left\{ N^{-2} \sum_u \frac{G(u)|F(u)|}{|G(u)|} e^{12\pi u \cdot x/N} \right\}$$

$$= 2N^{-2} \sum_{x \in S'} g'(x) \sum_u |F(u)| e^{12\pi u \cdot x/N} \frac{\partial}{\partial g(x_j)} \left\{ \frac{G(u)}{|G(u)|} \right\} . \qquad (F\text{-}6)$$

In order to evaluate the partial derivative in Eq. (F-6) we need expressions for $\frac{\partial G(u)}{\partial g(x_j)}$ and $\frac{\partial |G(u)|}{\partial g(x_j)}$:

$$\frac{\partial G(u)}{\partial g(x_j)} = \frac{\partial}{\partial g(x_j)} \sum_x g(x) e^{-12\pi u \cdot x/N}$$

$$= \sum_x e^{-12\pi u \cdot x/N} \frac{\partial g(x)}{\partial g(x_j)}$$

$$= e^{-12\pi u \cdot x_j/N} . \qquad (F\text{-}7)$$

The derivation for $\frac{\partial |G(u)|}{\partial g(x_j)}$ requires the result in Eq. (F-7):

$$\frac{\partial |G(u)|}{\partial g(x_j)} = \frac{1}{2|G(u)|} \frac{\partial}{\partial g(x_j)} |G(u)|^2$$

$$= \frac{1}{2|G(u)|} \left[G^*(u) \frac{\partial G(u)}{\partial g(x_j)} + C.C. \right]$$

$$= \frac{1}{2|G(u)|} \left[G^*(u) e^{-12\pi u \cdot x_j/N} + C.C. \right] \qquad (F\text{-}8)$$

where C.C. stands for complex conjugate of the explicit term. Using the results in (F-7) and (F-8) and with some algebraic manipulation the partial derivative in (F-6) becomes

$$\frac{\partial}{\partial g(x_j)} \frac{G(u)}{|G(u)|} = \frac{G^*(u)\, e^{-i2\pi u \cdot x_j/N}}{2|G(u)|\, G^*(u)} - \text{C.C.} \quad . \tag{F-9}$$

Substituting Eq. (F-9) back into Eq. (F-6) yields

$$\frac{\partial e_o^2}{\partial g(x_j)} = N^{-2} \sum_{x \in S'} g'(x) \sum_u |F(u)| \left[\frac{G^*(u)\, e^{-i2\pi u \cdot x_j/N}}{|G(u)|\, G^*(u)} - \text{C.C.} \right] e^{i2\pi u \cdot x/N} \quad . \tag{F-10}$$

By changing the order of summation we have

$$\frac{\partial e_o^2}{\partial g(x_j)} = N^{-2} \sum_u |F(u)| \left[\frac{G^*(u)\, e^{-i2\pi u \cdot x_j/N}}{|G(u)|\, G^*(u)} - \text{C.C.} \right] \sum_{x \in S'} g'(x)\, e^{i2\pi u \cdot x/N} \quad . \tag{F-11}$$

At this point it is convenient to define the characteristic function of the complement of the support S as follows

$$S'(x) = \begin{cases} 1 \ , \ x \in S' \\ \\ 0 \ , \ x \in S \end{cases} \quad . \tag{F-12}$$

The second summation in (F-11) may now be rewritten

$$\sum_{x \in S'} g'(x)\, e^{i2\pi u \cdot x/N} = \sum_x S'(x) g'(x)\, e^{i2\pi u \cdot x/N} \quad . \tag{F-13}$$

The error in the output, $g_e(x)$, consists of that component of the output that violates the support constraint:

$$g_e(x) = S'(x) g'(x) \quad . \tag{F-14}$$

The summation in Eq. (F-13) has the form of a forward DFT:

$$\sum_x g_e(x)\, e^{i2\pi u\cdot x/N} = \sum_x g_e(x)\, e^{-i2\pi(-u\cdot x)/N}$$

$$= G_e(-u) \qquad \text{(F-15)}$$

where $G_e(u)$ is the DFT of $g_e(x)$. The total partial derivative may now be written:

$$\frac{\partial e_o^2}{\partial g(x_j)} = N^{-2} \sum_u \frac{|F(u)|G_e(-u)}{|G(u)|G^*(u)}\left[G^*(u)\, e^{-i2\pi u\cdot x_j/N} - \text{C.C.}\right]$$

$$= N^{-2} \sum_u \frac{|F(u)|G_e(-u)}{|G(u)|}\, e^{-i2\pi u\cdot x_j/N} - N^{-2} \sum_u \frac{|F(u)|G_e(-u)G(u)}{|G(u)|G^*(u)}\, e^{i2\pi u\cdot x_j/N} \qquad \text{(F-16)}$$

Using Eq. (F-5) in the second term:

$$\frac{\partial e_o^2}{\partial g(x_j)} = N^{-2} \sum_u \frac{|F(u)|G_e(-u)}{|G(u)|}\, e^{-i2\pi u\cdot x_j/N} - N^{-2} \sum_u \frac{G'(u)G_e(-u)}{G^*(u)}\, e^{i2\pi u\cdot x_j/N} \qquad \text{(F-17)}$$

Remarkably both of these terms have the general form of DFTs. In order to combine both of these terms into a single inverse DFT we perform a sign change of variable on the Fourier vector in the first term. The net result is:

$$\frac{\partial e_o^2}{\partial g(x_j)} = N^{-2} \sum_u \left[\frac{|F(-u)|G_j(u)}{|G(-u)|} - \frac{G'(u)G_e(-u)}{G^*(j)}\right] e^{i2\pi u\cdot x_j/N} \qquad \text{(F-18)}$$

Finally, by appealing to the Hermitian property for Fourier transforms of real functions we may make the following substitutions:

$$|F(-u)| = |F(u)|$$

$$|G(-u)| = |G(u)|$$

$$G_e(-u) = G_e^*(u) \qquad \text{(F-19)}$$

to get the final result:

$$\frac{\partial e_o^2}{\partial g(x_j)} = N^{-2} \sum_u \left[\frac{|F(u)| G_e(u)}{|G(u)|} - \frac{G'(u) G_e^*(u)}{G^*(u)} \right] e^{12\pi u \cdot x_j / N} . \qquad \text{(F-20)}$$

Appendix G
$\nabla e_o^2(g(x))$ FOR COMPLEX OBJECTS

The derivation of the gradient of the $e_o^2(g(x))$ objective function for the case of complex objects closely follows that for real objects (Appendix F). Therefore, we just highlight this derivation, calling attention to significant differences. We begin by acknowledging that the complex case admits twice as many parameters; namely the real and imaginary parts of each input pixel. We denote the real and imaginary parts of the input function as follows:

$$g(x) = a(x) + i\, b(x). \tag{G-1}$$

We define the gradient to be:

$$\nabla e_o^2(g(x)) = \sum_{j=1}^{N^2} \frac{\partial e_o^2}{\partial a(x_j)} v_j^R + \frac{\partial e_o^2}{\partial b(x_j)} v_j^I \tag{G-2}$$

where v_j^R and v_j^I are orthogonal unit vectors associated with the parameters of the real and imaginary parts of the pixel at location x_j. The partial derivative of the objective function with respect to the real part of a pixel value may be written

$$\begin{aligned}
\frac{\partial e_o^2}{\partial a(x_j)} &= \frac{\partial}{\partial a(x_j)} \sum_{x \in S'} |g'(x)|^2 \\
&= \sum_{x \in S'} g'^*(x) \frac{\partial g'(x)}{\partial a(x_j)} + C.C. \\
&= \left[\sum_{x \in S'} g'^*(x) \frac{\partial}{\partial a(x_j)} \left\{ N^{-2} \sum_u G'(u)\, e^{i2\pi u \cdot x/N} \right\} \right] + C.C. \\
&= \left[N^{-2} \sum_{x \in S'} g'^*(x) \sum_u |F(u)|\, e^{i2\pi u \cdot x/N} \frac{\partial}{\partial a(x_j)} \left\{ \frac{G(u)}{|G(u)|} \right\} \right] + C.C.
\end{aligned} \tag{G-3}$$

The extra complex-conjugate term appears because the output functio g'(x), can assume complex values. The partial derivative with respec to the imaginary part is similarly found:

$$\frac{\partial e_o^2}{\partial b(x_j)} = \left[N^{-2} \sum_{x \in S'} g'^*(x) \sum_u |F(u)| \, e^{i2\pi u \cdot x/N} \frac{\partial}{\partial b(x_j)} \left\{ \frac{G(u)}{|G(u)|} \right\} \right] + C.C. \tag{G-4}$$

With simple algebraic manipulations the following useful identities may be verified:

$$\frac{\partial G(u)}{\partial a(x_j)} = e^{-i2\pi u \cdot x_j/N} \tag{G-5}$$

$$\frac{\partial G(u)}{\partial b(x_j)} = ie^{-i2\pi u \cdot x_j/N} \tag{G-6}$$

$$\frac{\partial |G(u)|}{\partial a(x_j)} = \frac{1}{2|G(u)|} \left[G^*(u) \, e^{-i2\pi u \cdot x_j/N} + C.C. \right] \tag{G-7}$$

$$\frac{\partial |G(u)|}{\partial b(x_j)} = \frac{i}{2|G(u)|} \left[G^*(u) \, e^{-i2\pi u \cdot x_j/N} - C.C. \right] . \tag{G-8}$$

With the aid of Eqs. (G-5) thru (G-8) we deduce

$$\frac{\partial}{\partial a(x_j)} \left\{ \frac{G(u)}{|G(u)|} \right\} = \frac{G^*(u) \, e^{-i2\pi u \cdot x_j/N} - C.C.}{2|G(u)|G^*(u)} \tag{G-9}$$

and

$$\frac{\partial}{\partial b(x_j)}\left\{\frac{G(u)}{|G(u)|}\right\} = \frac{iG^*(u)\, e^{-i2\pi u\cdot x_j/N} - C.C}{2|G(u)|G^*(u)} \quad . \tag{G-10}$$

When Eq. (G-9) is substituted into Eq. (G-3) and the order of summation is interchanged we get:

$$\frac{\partial e_o^2}{\partial a(x_j)} = \left[\frac{N^{-2}}{2}\sum_u |F(u)|\left\{\frac{G^*(u)\, e^{-i2\pi u\cdot x_j/N}}{|G(u)|G^*(u)} - C.C.\right\}\sum_x S'(x)g'^*(x)\, e^{i2\pi u\cdot x/N}\right] + C.C. \tag{G-11}$$

were S'(x) is the characteristic function of the set S', as before. Recall the object and Fourier-domain expressions for the error image:

$$g_e(x) = S'(x)g'(x) \tag{G-12}$$

$$G_e(u) = \sum_x S'(x)g'(x)e^{-i2\pi u\cdot x/N} \quad . \tag{G-13}$$

We may therefore write

$$G_e^*(u) = \sum_x S'(x)g'^*(x)e^{i2\pi u\cdot x/N} \tag{G-14}$$

which may be substituted into (G-11) to get:

$$\frac{\partial e_o^2}{\partial a(x_j)} = \left[\frac{N^{-2}}{2}\sum_u \frac{|F(u)|G_e^*(u)\left\{G^*(u)\, e^{-i2\pi u\cdot x_j/N} - C.C.\right\}}{|G(u)|G^*(u)}\right] + C.C. \tag{G-15}$$

By similar steps we find

$$\frac{\partial e_o^2}{\partial b(x_j)} = \left[\frac{N^{-2}}{2}\sum_u \frac{|F(u)|G_e^*(u)\left\{iG^*(u)\, e^{-i2\pi u\cdot x_j/N} - C.C.\right\}}{|G(u)|G^*(u)}\right] + C.C. \quad (G\text{-}16)$$

Explicitly writing out the complex-conjugate terms in Eqs. (G-15) and (G-16) and performing a few additional manipulations we produce the final result.

$$\frac{\partial e_o^2}{\partial a(x_j)} = \mathrm{Re}\left\{N^{-2}\sum_u \left[\frac{|F(-u)|G_e^*(-u)}{|G(-u)|} - \frac{G'(u)G_e^*(u)}{G^*(u)}\right] e^{i2\pi u\cdot x_j/N}\right\} \quad (G\text{-}17)$$

$$\frac{\partial e_o^2}{\partial b(x_j)} = -\mathrm{Im}\left\{N^{-2}\sum_u \left[\frac{|F(-u)|G_e^*(-u)}{|G(-u)|} + \frac{G'(u)G_e^*(u)}{G^*(u)}\right] e^{i2\pi u\cdot x_j/N}\right\} \quad (G\text{-}18)$$

It is gratifying that Eq. (G-17) is consistent with Eq. (F-18) which is the equivalent partial derivative for real objects only. It is worth mentioning that because the summation arguments in Eqs. (G-17) and (G-18) differ, an additional FFT is required in the computation of the gradient for complex objects. The total number of FFTs (forward or inverse) needed is increased to five.

Appendix H

DATA PROCESSING IN THE PHASE RETRIEVAL LABORATORY

The portable laboratory data acquisition and analysis system can acquire image data via a Fairchild CCD3000 camera; digitize, integrate, and process 512 by 512 image data via Imaging Technology video hardware; and execute algorithms requiring FFT and other computationally intensive operations (such as the iterative Fourier transform algorithm for phase-retrieval) using a Mercury ZIP 3232 16 Mflop array processor. The host computer is a Heurikon 68000-based system using the UNIX operating system. An outline of the functions of the experiment control software is given below.

1. Data acquisition

Digitize: Digitize and store image in Imaging Technology, Inc. (ITI) frame buffer with correction for calibrated nonuniformities.

Integrate: Digitize n images and sum in array processor (AP) with/without normalization.

2. Image display

Display: Display AP and hard disk images on ITI.

Notes: (1) Conversion from 32 bit to 8 bit data if desired

(2) Many options:

Display real, imaginary, magnitude, magnitude-squared, or phase

Apply bias and scale (as in y=ax+b)

Display absolute value

Magnify by 2,4,8,... (specify subimage to be displayed)

Sample to give 256x256 image and display in specified quadrant of ITI display (allows four images to be displayed simultaneously for comparison)

Display any size image in any location of display

(3) Values above and below the 8 bit range of the ITI are clipped at 0 and 255

Live/Memory: Toggle between displaying video incoming to ITI and data in ITI frame buffer.

3. Image algebra (all in AP)

Add: Add two images.

Subtract: Subtract two images.

Multiply: Multiply two images.

Divide: Divide two images with user definable result for divide by zero.

Scale: Add bias and scale image.

Threshold: Hard limit above and below.

Logic operations between binary images.

Magnitude: Find magnitude or magnitude-squared of an image.

Phase: Find phase of an image.

Convert: Change real image to/from complex image.

Print: Print values of specified small part of an image.

Statistics: Find mean, variance of image and magnitude-squared of an image.

Maxmin: Find max and min values of image.

Histogram: Compute histogram of image and display on ITI.

Convolve: Convolve image with a small specified convolution kernal (allows smoothing and other operations on data).

Interpolation: Interpolate from one sample spacing to another.

4. Create images (in AP) for test purposes

Zero: Zero fill an image.

Create: Place a rectangular, circular, or triangular region of specified complex value at a specified position in an image.

Aperture: Multiply image by a binary rectangular, circular, or triangular aperture located at a specified position.

Noise: Add zero-mean Gaussian noise with specified variance to an image (specifying seed so that same or different set of random numbers can be generated) (Also uniform and Poisson noise).

5. Iterative algorithm

   Setup: Allocate and load image domain, Fourier domain, image domain constraint, Fourier magnitude constraint, and buffer arrays in AP.

   Iterate: Iterate n times using specified form of iterative algorithm, computing and printing error measures.

   Display: Display intermediate results.

   Save: Save results.

6. Image error computation

   Error measure: Compute normalized root-mean-squared error of complex image or of image magnitude relative to reference object, taking into account intensity scaling and (for complex images) constant phase shift.

Appendix I

STOCHASTIC VS DETERMINISTIC APPROACHES TO PHASE RETRIEVAL

I.1 INTRODUCTION

When a SAR system operates under ideal conditions, the phase variation occurring from one received echo to the next is known. This knowledge is employed in the receiver by building a "matched filter"--matched to the amplitude and phase characteristics of the train of echoes that are processed during the coherent integration interval.

In some circumstances the phase variation from echo to echo is not completely known. This may occur, for example, due to uncompensated platform motion effects, unknown target accelerations, atmospheric effects, etc. In these cases, the lack of complete phase information can be modelled by considering the ideal situation (no phase error) to be perturbed by the presence of phase errors, $\epsilon_n(t)$, occurring on each (n^{th}) echo.

One approach for reducing the deleterious effects of uncompensated phase errors on the SAR imagery is to <u>remove</u> the phase in the recorded (pre-processed data), preserve the magnitude, and attach a pre-selected phase history using a priori information about the shape of objects in the final image. In this memo this approach is referred to as the <u>deterministic</u> approach to phase retrieval to distinguish it from another approach described next.

Usually the phase perturbations arise from a physical mechanism that can be modelled as a slowly varying (i.e., from pulse-to-pulse) stochastic process. If the statistics of that process are known, then the phase perturbations can be compensated out. To account for spatial

variations in the statistical model, however, it is necessary to regard the perturbation statistics as only partially known.

In the alternative phase retrieval scheme suggested here the phase errors are regarded as a zero mean Gaussian random process whose covariance function is known up to a finite number of unknown parameters. These unknown parameters are estimated during the actual system operation, using the received (phase-corrupted) data along with a priori knowledge about the dynamics of the physical process that actually caused the perturbations. Thus, the random process approach to phase retrieval exploits a different type of a priori information than the deterministic approach.

Depending upon the system being modelled and the scene being imaged, it is likely that these alternative phase retrieval methods will demonstrate different degrees of image enhancement. The deterministic approach should perform best when the variety of images being sought is limited to a sufficiently small set that each possible attached phase function can be considered. The random process approach ought to be favored when the scene is highly variable but where the physical causes of phase perturbation, itself, can be modelled as a (low dimensional) parameter estimation problem.

Below we outline the technical basis of stochastic phase retrieval. We begin by showing how the actual observables furnished by the sensor are affected by random phase errors. We then show that if the covariance function of the phase errors were known, then a <u>complete characterization</u> of the desired image could be formulated, just as in the deterministic case. It is then seen that the deterministic phase retrieval method attempts to estimate the proper phase for the actual waveform received from a target scene; while the stochastic approach attempts to estimate the proper phase of the waveform's <u>covariance function</u>.

I.2 TECHNICAL BASIS

If $\epsilon_n(t)$ has appreciable high-frequency energy, (i.e., if it changes rapidly in time) then even as one range sweep (an echo) arrives at the radar, we must account for the phase variation that occurs. It can be shown that in this case, the result of employing the processing described, e.g. in Brown [I.1] is to produce the "observables" on <u>one</u> (the n^{th}) echo (i.e., one range sweep):

$$Q(\omega, \underline{u}_n) = \int_{\underline{r}_o \epsilon R} \tilde{\sigma}(\underline{r}_o)\, e^{jk\nu_n(\underline{r}_o)}\, I_\epsilon(\underline{r}_o, \omega)\, dV(\underline{r}_o) \qquad (I\text{-}1)$$

where

$$I_\epsilon(\underline{r}_o, \omega) = \int_r \left[W[r + \nu_n(\underline{r}_o)] \exp[j\epsilon_n(r)]\right] e^{-j\omega r}\, dr$$

and where:

$\underline{r}_o$ denotes a vector drawn from the origin of a coordinate system centered somewhere on the target (scene) to each element of volume, $dV(\underline{r}_o)$, contributing to the range sweep.

R defines the region of space that contributes to the range sweep. (Normally R is determined by the receive system's antenna gain).

$\nu_n = \underline{u}_n \cdot \underline{r}_o$ is a scalar variable that depends on the vector $\underline{r}_o$ defined above and upon the <u>unit</u> vector, $\underline{u}_n$, defined along the radar's line-of-site during the round trip path traversed by a single transmitted pulse and its echo. (Platform/target motion is assumed "frozen" during the time required for the transmitted

pulse to traverse the interval of range contained in the range sweep).

$W(r) \triangleq A\left(\frac{2r}{c}\right)\exp\left\{j\phi\left(\frac{2r}{c}\right)\right\}$ is the complex envelope of the transmitted pulse, expressed as a function of $r = ct/2$.

$\epsilon_n(r)$ is the phase error occurring on the n^{th} echo.

Let us examine $I_\epsilon(\underline{r}_0,w)$ and show how our expression in (I-1) relates to Brown's corresponding result. (We will see that we obtain Brown's result if $\epsilon \equiv 0$, which corresponds to the ideal case of no phase errors.)

Using Brown's notation define

$$Y(w) = \int_{-\infty}^{\infty} W(r)\, e^{-jwr}\, dr \quad . \tag{I-2}$$

Also, let

$$h_\epsilon^{(n)}(r) \triangleq \exp[j\epsilon_n(r)]$$

and

$$H_\epsilon^{(n)}(w) \triangleq \int_{-\infty}^{\infty} h_\epsilon^{(n)}(r)\, e^{-jwr}\, dr \quad . \tag{I-3}$$

It then follows that

$$I_\epsilon(\underline{r}_o, \omega) = \left\{e^{+j\omega\nu}\, Y(\omega)\right\} * \left\{H_\epsilon^{(n)}(\omega)\right\} \qquad \text{(I-4)}$$

where the symbol * denotes convolution.

In the case where $\epsilon_n(t) \equiv 0$, $h_\epsilon^{(n)}(r) \equiv 1$ and $H_\epsilon^{(n)}(\omega) = \delta(\omega)$, the Dirac delta distribution. Thus, the result of the convolution is in this case:

$$I_o(\underline{r}_o, \omega) = e^{+j\omega\nu}\, Y(\omega) \quad . \qquad \text{(I-5)}$$

Inserting (I-5) into (I-1) produces Brown's result [his Eq. (6)]:

$$Q_o(\omega, \underline{u}_n) = Y(\omega) \int \tilde{\sigma}(\underline{r}_o) \exp\left\{j(k + \omega)\, \underline{u}_n \cdot \underline{r}_o\right\} dV(\underline{r}_o) \qquad \text{(I-6)}$$

As described in Brown, the next step is normally to employ a filter matched to $Y(\omega)$. This filtering accomplishes the pulse compression in range and one then has

$$\tilde{Q}_o(\omega, \underline{u}_n) = T(\omega) \int \tilde{\sigma}(\underline{r}_o) \exp\left\{j(k + \omega)\, \underline{u}_n \cdot \underline{r}_o\right\} dV(\underline{r}_o) \qquad \text{(I-7)}$$

where the ~ denotes the output of the range compression filter and $T(\omega) = K(\omega)\, Y(\omega)$.

From the above it follows that when phase errors are present, the desired Fourier transform relationship between $\tilde{\sigma}(x,y)$ and $\tilde{Q}(\omega,\underline{u}_n)$ <u>is not generally true</u>. This is significant in the interpretation of the phase retrieval problem because it is not accurate to regard the phase retrieval problem simply as a problem in inverting a Fourier transform in all cases.

It is true, however, that we may still regard the observables as the Fourier Transform of the desired scene if $\epsilon_n(r)$ is assumed to be constant during any one range sweep. This means $\epsilon_n(r)$ is not a function of r at all. Following a procedure just like the one used above, it then follows that for phase errors that vary only from pulse-to-pulse:

$$\tilde{Q}(\omega, \underline{u}_n) = \exp(j\epsilon_n)\, \tilde{Q}_o(\omega, \underline{u}_n) \qquad (I\text{-}8)$$

where $\tilde{Q}_o(\omega,\underline{u}_n)$ are the error-free observables. Hence, in this case the observables, $\tilde{Q}(\omega,\underline{u}_n)$, are a version of the error-free case that has been rotated in phase by the phase error ϵ_n. This rotation (by different values of ϵ_n) occurs on each received range sweep in our model.

The key observation at this point is that under the most general conditions we must impose some constraint on the sequence of values $(\epsilon_1, \epsilon_2, \ldots, \epsilon_n, \ldots \epsilon_N)$ from N echoes if there is any hope of forming a SAR image. If we assume that all the ϵ_n are random variables, mutually independent and uniformly distributed over $[0, 2\pi]$, then it is not possible to "coherently" sum the N azimuth echoes except in the impractical situation of trying all possible combinations of ϵ_n that could jointly occur over the N echoes. [If each ϵ_n is divided into 8 equal values (discretization of phase) there would be $(8)^{(N-1)}$ phase combinations to try. Even if only 10 echoes were processed for the SAR resolution desired, exhaustive search of all these possibilities requires more than 10^8 combinations of phase to be tested]. With the deterministic approach to phase retrieval, the constraints that are imposed on the $\{\epsilon_i\}$ arise from assumed knowledge of target shape. The stochastic approach to constraining the $\{\epsilon_i\}$ employs knowledge of the correlation function of the random errors in order to compensate for them.

A view held by some investigators of the phase retrieval problem is that the correlation between the errors might turn out to be essentially zero. This will depend upon the causative mechanism and upon various radar parameters. Admittedly if the errors <u>are</u> uncorrelated, then the stochastic technique discussed here will not offer any image quality improvement. (In fairness, it is possible to conjure up circumstances which will confound the deterministic approach also.) However, there are many circumstances in which the effect of the phase errors is to degrade rather than to destroy the image. In these cases, the fact that SAR images can be formed even though the data may be corrupted by phase errors, arises from dependence between the $\{\epsilon_n\}$. Thus, as successive echoes are received we can take advantage of this correlation. Note that this is not equivalent to eliminating the factor $\exp(j\epsilon_n)$ in (I-8) by taking envelopes. Nor is it true that only the envelope is observable. The set of observables available to us is the collection of complex values

$$\tilde{Q}(\omega,\underline{u}_1),\ \tilde{Q}(\omega,\underline{u}_2),\ \ldots\ \tilde{Q}(\omega,\underline{u}_n),\ \ldots\ \tilde{Q}(\omega,\underline{u}_N)$$

from N received echoes, preserving their individual phases, even though they contain the errors. (Alternatively we could work directly with the data in the time domain.)

It turns out that <u>relative</u> phase is the only thing that matters in forming the coherent sum of the N samples. Hence, for N echoes, there are (N-1) unknown phases. There are several ways of approaching the problem of estimating this (N-1)-dimensional phase vector, but they all rely on employing some kind of structural constraint on the variation of ϵ_n from one echo to the next, as they should, because without such a structure we have already stated that exhaustive search is all that is left.

STOCHASTIC APPROACH

One approach is to regard the ϵ_n as samples of an N-dimensional stochastic process that characterizes the phase errors over the time history involved in the coherent integration process. For example, suppose the collection of $\tilde{Q}$ observables occurring over N echoes can be modelled as samples from a Gaussian random process. Then the N echoes can be completely described statistically using the mean and covariance function of the underlying process.

For the moment assume that the mean is known. Then it is no problem if we also assume that the mean is zero. In this case all information is embodied in the covariance function of the underlying process. The random process we are speaking of generates an N-dimensional joint Gaussian probability density function which requires specifying the N X N covariance matrix of the $\{\epsilon_n\}$, [actually of $\exp(j\epsilon_n)$]. This correlation matrix results from sampling a continuous process, and we can imagine that there is an underlying covariance function

$$\rho(t,s) = \langle \exp(j\epsilon_t) \exp(-j\epsilon_s) \rangle = \langle \exp[j(\epsilon_t - \epsilon_s)] \rangle \quad . \qquad \text{(I-9)}$$

In general $\rho(t,s)$ is complex and it may be separated into its magnitude and phase

$$\rho(t,s) = |\rho(t,s)| \, e^{j\theta(t,s)} \quad . \qquad \text{(I-10)}$$

If the observables can be modelled as jointly Gaussian (i.e., as samples from a Gaussian random process), then it is possible to perform a series expansion for any sample function and the resulting expansion is analagous to the representation of a <u>deterministic</u> time waveform--but there <u>are</u> some important differences.

Suppose we consider the problem of representing a one-dimensional function, $v(t)$, over an interval $[-T/2, T/2]$. If $v(t)$ is a deterministic waveform, it has a Fourier series representation:

$$v_N(t) = \sum_{k=0}^{N} v_k \exp(j\omega_k t) \; ; \; \omega_k = \frac{2\pi k}{T} \qquad \text{(I-11a)}$$

where

$$v_k = \int_{-T/2}^{T/2} v(t) \exp(-j\omega_k t)\, dt = \int_{-T/2}^{T/2} |v(t)|\, e^{j\alpha(t)} \exp(-j\omega_k t)\, dt \; . \qquad \text{(I-11b)}$$

Note that in general the coefficients, v_k, are complex and they are functions of the phase function, $\alpha(t)$, of $v(t)$. Thus, we should more carefully write $v_N(t) = v_N(t;\alpha(t))$. The property of the expansion in (I-11) is that the series converges everywhere to $v(t)$.

$$\lim_{N\to\infty} |v_N(t)\ \alpha(t)) - v(t)|^2 = 0 \qquad \text{(I-12)}$$

If the phase, $\alpha(t)$, in (I-11b) is unknown, then we cannot form the series expansion unless some estimate, $\hat{\alpha}(t)$, of $\alpha(t)$ is available. Leaving aside for the moment just how $\hat{\alpha}(t)$ can be formulated, its utility is based on the requirement [for the deterministic $v(t)$] that

$$\lim_{N\to\infty} |v_N(t;\, \hat{\alpha}(t)) - v(t)|^2 = 0 \; . \qquad \text{(I-13)}$$

For brevity we'll denote $v_N(t;\hat{\alpha}(t)) \equiv \hat{v}_N(t)$. From (I-11b) and (I-13) it is seen that for the deterministic case, the problem of representing $v(t)$ accurately is equivalent to knowing $|v(t)|$ and $\alpha(t)$. In this case, there is no problem to solve (except to take more terms in

the series) if we assume that both $v(t)$ and $\alpha(t)$ are known. The deterministic phase retrieval problem consists of assuming that $\alpha(t)$ is unknown and searching for conditions under which $\alpha(t)$ can be uniquely inferred from $|v(t)|$. In the one-dimensional deterministic signal case there are no generally useful conditions under which this can be done. To overcome the lack of constraints between $\alpha(t)$ and $|v(t)|$ in the one-dimensional deterministic case, one possibility is to invoke relationships between $\alpha(t)$ and $|v(t)|$ that arise from exploiting object shape information. When such information is at hand it is certainly worth exploiting. Another possible approach is to regard $\alpha(t)$ as a known function of a finite (and small) number of unknown deterministic parameters. A third alternative, and the one explored below, exploits a different source of auxiliary information. So let's see what happens if we regard the function $v(t)$ as a stochastic process.

When $v(t)$ is a sample function from a Gaussian random process, it turns out that there is an expansion [Karhunen-Loeve expansion] for $v(t)$ that corresponds to the Fourier series in (I-11). The desired expansion is described as follows:

$$v_N(t) = \sum_{k=0}^{N} v_k \, \Psi_k(t) \qquad \text{(I-14a)}$$

where

$$v_k = \int_{-T/2}^{T/2} v(t) \, \Psi_k^*(t) \, dt \qquad \text{(I-14b)}$$

and where $\{\Psi_k(t)\}_{k=1}^{\infty}$ is the set of eigenfunctions with eigenvalues $\{\lambda_k\}_{k=1}^{\infty}$ that satisfy the following integral equation (the eigenfunctions are ortho-normal):

$$\lambda_k \Psi_k(t) = \int_{-T/2}^{T/2} \rho(t,s)\, \Psi_k(s)\, ds \quad . \qquad \text{(I-14c)}$$

Note that the $\{\Psi_k(t)\}$ are completely determined from knowledge of $\rho(t,s)$.

It can be shown that in the special case of "stationarity" where $\rho(t,s) = \rho(t-s)$ and where T is large compared to the reciprocal bandwidth of the power spectrum, $S_\rho(w) = \mathcal{F}\{\rho(t,s)\}$, the solution to (I-14c) is the set of functions $\Psi_k(t) = \exp(j\omega_k t)$ and the λ_k are simply the values of $S_\rho(\omega_k)$. Thus, in this case of stationarity and large observation interval, the expressions in (I-11a, b) and (I-14a, b) are algebraically identical. Nonetheless the interpretation of the two results is different and (I-14a, b) is more general. For one thing, in the stochastic case, the most we can hope for is that

$$\langle \lim_{N\to\infty} |v_N(t) - v(t)|^2 \rangle = 0 \qquad \text{(I-15)}$$

where $\langle \ldots \rangle$ denotes expectation (averaging) over the ensemble of functions generated by the underlying random process. Moreover, (I-15) is true only if the function $\rho(t,s)$ in (I-14c) is positive definite. (This means it has no zero eigenvalues). This is not a difficult requirement to satisfy in practice. For example, the result of passing white noise through any linear, stable, realizable filter will yield a random process whose covariance function is strictly positive definite.

It should be pointed out that the convergence of $v_N(t)$ to $v(t)$ in (I-15) is in general weaker than the convergence in (I-12). However, if $v(t)$ is a sample function from a Gaussian random process, then the two are essentially the same (except on a set of measure zero!).

Finally, we hasten to add that the expansion in (I-14) does not require stationarity, so it is a more general treatment than Fourier series alone would permit. Since the Fourier series approaches the Fourier transform as $T \to \infty$ (in the deterministic case), we can expect conditions we may be looking for in the deterministic case may have no counterpart in the stochastic case (unless the random process is stationary, of course).

I.3 SUMMARY

We now know that to represent a Gaussian random process uniquely over $[-T/2, T/2]$ we need (assuming zero mean)

(1) $\rho(t,s) = \langle v(t)v^*(s) \rangle$
(2) Positive definiteness of $\rho(t,s)$
(3) Any sample function, $v(t)$, observed over $[-T/2, T/2]$.

Thus, if $\rho(t,s)$ is completely known, there is no reconstruction problem remaining, and this is true even if the sample function, $v(t)$, contains a phase function that is perturbed by a random unknown component. Whereas in the deterministic case the phase of the underlying sample had to be considered unknown to still have a problem to solve, in the random process formulation a problem of interest occurs as long as the covariance function is regarded as incompletely specified. This is an important conceptual difference between the two problems.

The physical significance of the above is that the deterministic situation seeks constraints on the exact values (for uniqueness) that a particular time waveform can have, whereas in the random model the constraints involve the behavior (correlation) between different sample functions, drawn from a statistically similar ensemble.

The stochastic version of phase retrieval

Based on the above we may consider the following non-trivial phase retrieval problem for one-dimensional data. (The notions apply equally well for two-dimensional data, but unlike the deterministic approach, we do not encounter an unsolvable one-dimensional problem so we illustrate the ideas only for that case now.)

Let $v(t) = a(t)e^{j\theta(t)}$ be a sample function of a random process observed over $[-T/2, T/2]$. Assume $\langle v(t) \rangle = 0$ and

$$\rho(t,s) = \langle v(t)\, v^*(s) \rangle = \langle a(t)\, a^*(s) \rangle \langle e^{j[\theta(t) - \theta(s)]} \rangle \quad \text{(I-16)}$$

where the "envelope" and "phase" random processes $a(t)$ and $\theta(t)$, have respectively been assumed to be mutually independent.

Hence,

$$\rho(t,s) = \rho_a(t,s)\ \rho_\theta(t,s) \quad \text{(I-17)}$$

where ρ_a is the covariance of the amplitude and ρ_θ is the covariance of the phase.

From (I-14), if ρ_a and ρ_θ were known, we could (in principle) solve (I-14c) for $\{\Psi_k(t)\}$ and $\{\lambda_k\}$ and then obtain our desired expansion. We can, however, regard ρ_θ as incompletely known, and consider the estimation of this function as the phase-retrieval problem.

Estimating ρ_θ

Without some kind of parameterization of the problem it is hopeless to attempt to obtain a useful estimate of ρ_θ from a limited quantity of

data. However, in many cases it is reasonable to model $\theta(t)$ as the result of certain dynamic processes that must satisfy laws of motion.

For example, if $\theta(t)$ is the residual phase error resulting from uncompensated target/platform motion effects then

$$\theta(t) = \frac{2\pi r(t)}{\lambda} = \frac{2\pi}{\lambda}\left\{\dot{r}(t_o) + \dot{r}(t_o)\ (t - t_o) + \frac{1}{2}\ddot{r}(t_o)\ (t - t_o)^2\right\} \tag{I-18}$$

where $r(t_o)$, $\dot{r}(t_o)$, $\ddot{r}(t_o)$ are random variables, fixed during one record. Then estimating ρ_θ is equivalent to estimating these constants based on the data collected over an interval $[-T/2, T/2]$. This can be interpreted, in some cases, as equivalent to estimating certain parameters of the observed processes' power spectrum, such as its bandwidth. It is possible to employ the rigors of statistical estimation theory to derive suitable estimates of the required parameters of $\theta(t)$ using the corrupted observable data, and it is also possible to derive expressions for the accuracy attainable in estimating ρ_θ.

The theory required to develop the required estimates exists and does not require any conceptual or mathematical breakthroughs--but it is different from the deterministic phase retrieval problem.

It has been demonstrated that the problem of estimating the phase in a phase-corrupted version of received radar data can be viewed as a problem of estimating a portion of the received data's correlation function. This viewpoint is capable of producing useful phase estimates even on one-dimensional (time-series) data without requiring any information on target shape. Instead, a priori information about the physics of the dynamic processes that cause the undesired phase disturbance is exploited. The ideas suggested here are therefore

adjuncts to the deterministic phase retrieval methods currently being pursued. Use of both approaches should further enhance phase retrieval performance.

Reference

I.1 W.M. Brown, "Walker Model for Radar Sensing of Rigid Target Fields," IEEE Trans. AES AES-16, 104-107 (1980).